L'apport des informations visuelles des gestes oro-faciaux dans le traitement phonologique des phonèmes natifs et non-natifs : approches comportementale, neurophysiologique

Sabine Burfin

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Sciences Cognitives, Psychologie et Neurocognition**

Arrêté ministériel : 7 août 2006

Présentée par

Sabine BURFIN

Thèse dirigée par **Sonia KANDEL**

préparée au sein du **Laboratoire de Psychologie et NeuroCognition - CNRS UMR 5105**
dans **l'École Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement**

L'apport des informations visuelles des gestes oro-faciaux dans le traitement phonologique des phonèmes natifs et non-natifs : approches comportementale et neurophysiologique

Thèse soutenue publiquement le **3 Février 2015**,
devant le jury composé de :

Madame, Fanny, MEUNIER
DR CNRS, Bron, Rapporteur
Monsieur, Noël, NGUYEN
Professeur, Université Aix-Marseille, Rapporteur
Monsieur, Jean-Luc, SCHWARTZ
DR CNRS, Grenoble, Président de jury et Examinateur
Madame, Sonia, KANDEL
Professeur, Université de Grenoble, Directrice de thèse

Université Joseph Fourier / Université Pierre Mendès France / Université Stendhal / Université de Savoie / Grenoble INP

A mes parents

A leur parents

A ma sœur

RESUME

En situation de perception audiovisuelle de la parole, comme lors des conversations face-à-face, nous pouvons tirer partie des informations visuelles fournies par les mouvements oro-faciaux du locuteur. Ceci améliore l'intelligibilité du discours. L'objectif de ce travail était de déterminer si ce « bénéfice audiovisuel » permet de mieux identifier les phonèmes qui n'existent pas dans notre langue. Nos résultats révèlent que l'utilisation de l'information visuelle permet de surmonter les difficultés posées par la surdité phonologique dont nous sommes victimes lors d'une présentation auditive seule (Etude 1). Une étude EEG indique que l'apport des informations visuelles au processus d'identification de phonèmes non natifs pourrait être dû à une modulation précoce des traitements effectués par le cortex auditif primaire (Etude 2). En présentation audiovisuelle les phonèmes non natifs donnent lieu à une P50, ce qui n'est pas observé pour les phonèmes natifs. Il semblerait également que l'expérience linguistique affecte l'utilisation des informations visuelles puisque des bilingues précoces semblent moins aptes à exploiter ces indices pour distinguer des phonèmes qui ne leur sont pas familiers (Etude 3). Enfin, l'étude de l'identification de consonnes plosives natives avec une tâche de dévoilement progressif nous a permis d'évaluer la contribution conjointe et séparée des informations auditives et visuelles (Etude 4). Nous avons observé que l'apport de la modalité visuelle n'est pas systématique et que la prédictibilité de l'identité du phonème dépend de la saillance visuelle des mouvements articulatoires du locuteur.

Mots-clés : perception audiovisuelle de la parole, phonèmes natifs et non natifs, bilinguisme, neurophysiologie.

ABSTRACT

During audiovisual speech perception, like in face-to-face conversations, we can take advantage of the visual information conveyed by the speaker's oro-facial gestures. This enhances the intelligibility of the utterance. The aim of this work was to determine whether this "audiovisual benefit" can improve the identification of phonemes that do not exist in our mother tongue. Our results revealed that the visual information contributes to overcome the phonological deafness phenomenon we experience in an audio only situation (Study 1). An ERP study indicates that this benefit could be due to the modulation of early processing in the primary auditory cortex (Study 2). The audiovisual presentation of non native phonemes generates a P50 that is not observed for native phonemes. The linguistic background affects the way we use visual information. Early bilinguals take less advantage of the visual cues during the processing of unfamiliar phonemes (Study 3). We examined the identification processes of native plosive consonants with a gating paradigm to evaluate the differential contribution of auditory and visual cues across time (Study 4). We observed that the audiovisual benefit is not systematic. Phoneme predictability depends on the visual saliency of the articulatory movements of the speaker.

Key words: audiovisual speech perception, native and non native phonemes, bilingualism, neurophysiology

FINANCEMENT

Cette thèse a été financée par une allocation de recherche de la région Rhône-Alpes, ARC8 "Industrialisation et Sciences de gouvernement". Elle a été réalisée au sein du Laboratoire de Psychologie et de NeuroCognition (CNRS UMR 5105) et de l'Ecole Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement.
Grâce au programme Explora'Doc, j'ai pu réaliser un séjour de trois mois au sein du Center of Brain and Cognition (Universitat Pompeu Fabra) à Barcelone, Espagne.
Je remercie donc doublement la région Rhône-Alpes (Rhône-Alpes).

Je tiens dans un premier temps à remercier ma directrice de thèse, Sonia Kandel, qui m'a fait confiance depuis ma seconde année de Master et qui a bien voulu m'accompagner dans cette aventure qu'est la thèse. Elle m'a toujours poussé vers l'avant, moi, la petite « gothique » repliée sur moi-même et m'a donné l'occasion de surmonter beaucoup de mes peurs tout au long de ces quatre ans. Je ne l'en remercierai jamais assez pour ça. Elle a contribué à apporter un peu de couleur dans ma vie (et mes habits), m'a permis de voyager, d'enseigner, de rencontrer de nombreuses personnes, bref, elle m'a ouverte au monde.

Je tiens dans un deuxième temps à témoigner ma reconnaissance au LPNC, et particulièrement à Monica Bacciu qui a veillé à ce que cette thèse se passe dans les meilleures conditions et m'a toujours apporté son soutien. Pour ces mêmes raisons, je remercie également Claire Leroy et Guylaine Omnes. Claire, je sais que tu t'es arrachée les cheveux de nombreuses fois à cause de moi, depuis la signature de mon contrat jusqu'à la fin de celui-ci (et même après), mais je tenais à saluer ton implication et ta ténacité qui m'ont sorti de la panade de nombreuses fois.

Mes pensées vont ensuite aux techniciens, chercheurs et chercheurs en devenir de mon laboratoire d'accueil mais aussi vers ceux du Gipsa-Lab qui ont fait de cette thèse une aventure des plus agréable et ont largement facilité mon travail que ce soit par un support technique, par des collaborations ou simplement par leur présence et leur amitié.

Dans la catégorie "*Support technique (mais pas que)*", les nominés sont : Eric Guinet, le roi d'Eprime et de la synchronisation, qui a toujours prêté une oreille attentive à mes (nombreux) problèmes techniques, Christophe Savariaux, le maître des enregistrements caméra rapide, qui a lui aussi fait preuve de beaucoup de patience pour m'initier aux subtilités de la linguistique, et sans qui mon niveau de maîtrise de Praat et Matlab serait resté au plus bas. Je remercie également Sylvain Harquel, le fou de l'électrode, dont la compagnie fût largement appréciée, voire salvatrice, durant les longues heures de passations. Je sais qu'il n'en est pas sorti indemne puisqu'aujourd'hui, un frisson l'envahit dès qu'il entend le mot « fait» ou encore « ça ».

Dans la catégorie "*Collaborations*", les nominés sont : Olivier Pascalis, qui m'a donné l'opportunité de m'intéresser à la perception, non plus des phonèmes, mais des visages. Il fût le meilleur des guides que l'on peut avoir dans le monde du « Other-race-effect ». Je nommerais également Marc Sato et Aurélie Campagne, qui ont été mon sésame d'entrée pour le monde électrisant de l'EEG. Je remercie également Laurent Vercueil pour son implication et son aide dans ce projet.

Dans la sous-catégorie "*Collaborations Internationales*", je tiens à citer Salvador Soto-Faraco et Albert Costa du CBC (Barcelone) qui m'ont donné l'opportunité de travailler à leurs côtés et qui se sont investis dans nos recherches. J'en profite pour remercier leurs étudiantes, Elisa Ruiz-Tada et Clara Martin pour le travail de passation qu'elles ont effectué à Barcelone, et particulièrement Carolina Sanchez pour son aide lors des passations mais également pour son soutien. Sans le savoir, elle m'a évité de nombreuses prises de tête, voire quelques crises de nerf. Dans ce sens et comme nous sommes dans les remerciements transpyrénéens, je souhaite exprimer toute ma gratitude à Salvador pour m'avoir accueilli durant trois mois dans son équipe au CBC. Cette expérience fut des plus enrichissantes et m'a beaucoup apporté, aussi

bien au niveau de la recherche qu'au niveau humain. Les qualités de chercheurs de Salvador n'étant plus à montrer, je soulignerai plutôt sa gentillesse et sa patience, qui m'ont permis d'appréhender ce séjour d'une façon plus que positive. Durant ces quelques mois, j'ai apprécié faire de nouvelles rencontres (je pense notamment à Silvia et Cristina) mais également me faire de nouveaux amis. J'ai une pensée très émue pour mes colocs, Alvaro Canario, Alvaro Valenciano et Marcito qui ont rendu ce séjour inoubliable. Ils ont été ma famille pendant trois mois et je sais que ce voyage n'aurait pas été aussi fou sans eux.

Enfin, dans la catégorie "*Alliés*" sont nominés tous ceux qui, à Grenoble, m'ont permis de mener cette thèse à terme par leur soutien et leur amitié au quotidien. Mes pensées vont d'abord, sans toutefois vouloir hiérarchiser ces remerciements, à Mel et Fab, que je connais maintenant depuis plus de huit ans. J'ai conscience que sans eux, cette thèse (ainsi que ma santé mentale) ne serait pas ce qu'elle est aujourd'hui. Je les remercie chaleureusement pour avoir toujours prêté une oreille attentive à mes questions, pour leur soutien, bref, pour leur simple présence dans les bons comme dans les mauvais moments. Vous êtes tous les deux des personnes de grande valeur, comme il y en a trop peu, et comptez parmi les plus gentilles que j'aurais l'occasion de rencontrer. Je pense aussi à ma petite Louise, qui fût ma copine de pause à moi (mais pas que), puisque les autres ont fini par devenir des "rats de laboratoire" qui ne voyaient plus la lumière du jour, et ce, à raison. Ton sens de l'humour et ta personnalité font de toi quelqu'un de remarquable, surtout ne change rien. Enfin, même s'il serait long de citer tous les nominés de cette catégorie, je me risque tout de même à remercier Ben et Ben, Marcé, Solène, Anne et Anne, Lysianne, Pauline, Lorie, Fanny, Marie, qui ont largement détendu les conversations et qui illuminaient, par leur bonne humeur et leur brin de folie, la vie en 222 bis.

Je prends également le temps de remercier toutes les personnes qui ont bien voulu participer à mes études, car c'est avant tout grâce à eux que ces travaux, et donc cette thèse, ont pu prendre forme.

Mes derniers remerciements iront à mes proches. En premier lieu, je remercie Guillaume, mon chercheur à moi, qui, même s'il est lui-même en thèse, a toujours débloqué du temps, beaucoup de temps même, pour que ces épreuves que sont la rédaction et la soutenance me semblent surmontables. Bien que son soutien moral ait été sans faille, c'est son aide en tant que « professionnel » que je citerai ici. J'ai eu la chance de bénéficier des critiques et conseils d'un passionné de parole, ce qui a largement contribué à améliorer cette thèse. Je le remercie donc pour sa patience, son temps (précieux) et simplement pour m'avoir accompagné et supporté pendant cette aventure qui a été bonifiée par sa présence.

Je voudrais enfin remercier ma famille, simplement pour avoir fait de moi ce que je suis aujourd'hui. Je leur dédie cette thèse qui, même si elle est le fruit ne mes efforts, n'a été possible que parce qu'ils m'ont donné, à chaque grande étape de ma vie, ce coup de pousse qui permet de continuer d'avancer. C'est parce qu'ils ont cru en moi que j'écris ces quelques lignes aujourd'hui. Un merci particulier à ma mère, mon grand-père et surtout ma sœur, qui, par leurs nombreux ravitaillements, m'ont évité une mort par malnutrition lors de ma rédaction et à mon papa, pour avoir supporté mes sautes d'humeur.

Merci à tous du fond du cœur, et même s'il y a des noms que je n'ai pas cité, parce qu'il ne serait pas de bon ton de faire des remerciements sans fin, je n'oublierai aucune des personnes que j'ai eu l'occasion de côtoyer durant ces trois années.

TABLE DES MATIERES

LISTE DES ABREVIATION

NB : Les abréviations en italique sont liées aux statistiques présentées dans ce manuscrit.

A Auditif, Auditive, Audio

AV Audiovisuel, Audiovisuelle

V Visuel, Visuel

ANOVA Analyse de variance (analysis of variance)

CV Consonne-Voyelle CVC Consonne-Voyelle-Consonne

CVCV Consonne-Voyelle-Consonne-Voyelle

dB Décibels

DC Détection correcte

EEG/MEG électro/magnétoencéphalogramme

F Indice statistique suivant la loi de Fisher

GA Gyrus Angulaire

GSM Gyrus Supra Marginal

GTI Gyrus Temporal Inférieur

GTM Gyrus Temporal Moyen

GTS Gyrus Temporal Supérieur

ISI : Intervalle Inter Stimuli

IRMf Imagerie par Résonance Magnétique fonctionnelle

ISI Intervalle Inter Stimuli

M Moyenne

N Natif, qui exisdte dans la langue maternelle

NN Non natif, qui n'existe pas dans la langue maternelle

NC Natif contrôle

p Probabilité associée à un indice statistique

RSB Rapport Signal Sur Bruit

STS Sillon Temporal Supérieur

t Indice statistique suivant la loi de Student

TMS Stimulation Magnétique Transcrânienne

Trad. Traduction

VC Voyelle-Consonne

VOT Voiced Onset Time

1.1.1.1 PREAMBULE

Le langage est ce qui nous distingue des *pecudesmutae* (i.e., animaux privés de parole). Il est l'attribution de symboles abstraits ou de sons (i.e., un signifiant) à des objets au sens large du terme (i.e., un signifié) pour permettre de communiquer des besoins, envies, états ou émotions (cf. Saussure, 1916). C'est un des grands pas de l'histoire qui a amené l'homme à se construire en tant qu'être civilisé. Le langage est l'objet de nombreuses questions depuis les grecs qui recherchaient déjà le « langage originel ». Depuis cette époque, la recherche expérimentale a permis de mieux comprendre les mécanismes de perception inscrits dans la compréhension de la parole. Son étude a dans un premier temps consisté à la considérer comme purement unimodale, ne s'intéressant qu'à son versant auditif. Cependant, l'Homme, en tant qu'animal social, évolue en interaction avec ses semblables. La plupart des situations de communication mettent donc en scène des individus face-à-face. Cela a amené, après plusieurs décennies de recherche, à considérer la parole, non plus comme reposant uniquement sur son aspect auditif, mais comme un événement *multimodal* à part entière. Ainsi, dans la majeure partie de ces interactions, l'Homme pourra tirer partie des informations auditives, mais également se reposer sur les indices visuels, notamment en utilisant les informations visuelles fournies par la mise en mouvement des articulateurs visibles de l'interlocuteur (i.e., lèvres, mâchoire, langue, etc.). L'utilisation conjointe et automatique des deux canaux permet de rendre le processus de communication aussi fluide que possible et de faciliter, dans des environnements bruités, la compréhension des unités de langage (i.e., phonèmes). Ces indices visuels se révèlent également importants lors de la perception de *phonèmes d'autres langues*. En effet, les phonèmes non natifs sont parfois confondus avec des phonèmes natifs par l'auditeur, les informations visuelles permettraient dans certains cas de mieux les identifier et donc d'améliorer la compréhension du discours.

L'objectif de cette thèse consiste à examiner si l'individu "tout-venant" est capable d'exploiter les informations visuelles dans le cadre de la perception de langues étrangères. En d'autres termes, nous évaluerons la contribution spécifique de la gestualité oro-faciale verbale (i.e., non émotionnelle) au processus d'identification des phonèmes qui n'existent pas dans notre langue maternelle. Pour cela, nous effectuerons dans un premier temps une revue de littérature concernant la perception auditive des phonèmes natifs pour nous diriger dans un second temps vers la perception auditive des phonèmes non natifs (Chapitre 1). Par la suite, nous présenterons les travaux portant sur la perception audiovisuelle de la parole native. Ils

seront la base de la réflexion sur les avantages perceptifs que peut amener la lecture labiale dans la perception audiovisuelle des phonèmes non natifs (Chapitre 2).

Dans la deuxième partie du manuscrit, nous présenterons le travail expérimental qui a été réalisé pour répondre à ces questionnements. Nous examinons le rôle de l'information visuelle dans le processus de perception de phonèmes en modalité auditive et audiovisuelle. Nous interrogerons dans un premier temps la capacité de monolingues francophones et hispanophones à utiliser des informations visuelles fournies par un locuteur prononçant des phonèmes inconnus afin de désambigüiser un contraste phonologique non natif (Etude 1). Par la suite, nous examinerons quelles sont les sous-bassements neurophysiologiques qui pourraient, de manière précoce, être impliqués lors de la perception audiovisuelle de phonèmes inconnus (Etude 2). Puis, nous nous intéresserons à une population particulière, les individus bilingues (Etude 3), pour étudier l'impact de la familiarisation précoce à plusieurs langues et donc à plusieurs codes labiaux sur les capacités à utiliser les informations visuelles lors de la perception de contrastes phonologiques non natifs. Nous clôturerons cette partie expérimentale en présentant les résultats préliminaires d'une étude portant sur le processus d'identification de phonème natifs lors d'une tâche de *gating on-line* (Etude 4). Le but était de comprendre comment les systèmes auditif et visuel décodent les informations fournies par les gestes oro-faciaux du locuteur en fonction du temps. C'est au septième et dernier chapitre que nous discuterons l'ensemble des résultats expérimentaux. Dans cette discussion générale nous montrons quels sont les apports de notre travail à la compréhension des processus perceptifs que l'on met en place lorsque nous sommes dans des situations de communication dans une langue qui n'est pas la nôtre. Nous exposons également les limites et perspectives de ces études.

CHAPITRE 1

LA PERCEPTION AUDITIVE DE LA PAROLE

« Language is the most massive and inclusive art we know,
a mountainous and anonymous work of unconscious generations. »

— Edward Sapir

Le langage est universel. Alors que nous sommes les seuls êtres vivants de la planète à posséder un langage articulé, et malgré la complexité de cette tâche, nous l'effectuons sans même y penser, de la manière la plus naturelle qui soit. L'investigation scientifique de ce domaine a fait couler beaucoup d'encre et a permis de poser des bases solides concernant les mécanismes recrutés pour nous comprendre les uns les autres. Cette compréhension a pour première étape la perception auditive des sons de la langue. Nous allons dans ce premier chapitre faire un inventaire succinct des travaux qui ont permis de définir les notions clés de la perception de la parole et de construire les modèles qui la décrivent. Ceci nous permettra ne nous pencher par la suite sur le challenge posé par la perception des phonèmes qui n'existent pas dans notre langue maternelle.

1.2 PERCEPTION DE LA PAROLE NATIVE

1.2.1 GENERALITES

La perception de la parole, même si elle nous parait simple au premier abord, parce que nous n'avons en effet pas besoin de faire d'effort particulier pour comprendre quelqu'un qui nous parle, relève de nombreux processus cognitifs, qui nous permettent d'extraire des unités discrètes d'un signal acoustique continu. En amont de la reconnaissance d'un mot, de nombreuses étapes sont nécessaires pour interpréter les variations du signal acoustique, notamment au travers de trois niveaux d'analyse : *acoustique*, *phonétique* et *phonologique*. La littérature scientifique s'accorde sur le fait que le traitement de la parole commence avec l'analyse du signal acoustique qui permet de dégager les traits acoustiques[1] et phonétiques[2] qui le constituent. Ces traits sont liés aux modifications de configuration des articulateurs lors de la production de la parole. Ils sont par la suite intégrés pour former les représentations phonémiques ou phonèmes. Chaque phonème est donc constitué d'un certain nombre de traits phonétiques.

Les consonnes se distinguent toutes entre elles sur au moins un des traits suivants : le voisement[3], la place d'articulation[4], le mode d'articulation[5] et la nasalité. La consonne /p/ sera par exemple identifiable grâce à la somme de traits phonétiques qui lui sont spécifiques (i.e., non voisée + bilabiale + occlusion + orale) et qui permettront de la distinguer de /b/ (i.e., voisée + bilabiale + occlusion + orale). Les voyelles sont décrites en fonction de la nasalité[6], du degré d'ouverture du conduit vocal, de la position avant ou arrière de la langue et de la labialité. Un phonème est la plus petite unité distinctive d'une langue qui permet une distinction sémantique entre deux mots (Saussure, 1916). Chaque langue utilise un nombre fini de phonèmes : environ 30 par langue. Par exemple, en français, le mot « pain » est composé

[1] Fréquence, amplitude, formant, etc

[2] Voisement, ouverture, ou tout trait exploité dans une langue donnée

[3] Un son est dit « voisé » si sa production s'accompagne d'une vibration des cordes vocales. Par exemple /s/ est non voisé alors que /z/ est voisé.

[4] La place ou lieu d'articulation est l'endroit où le passage de l'air est plus étroit, suite à un resserrement des lèvres ou à un rapprochement de la langue vers une partie du palais ou du pharynx (Hardison, 1999, 2003).

[5] La manière ou mode d'articulation décrit l'interaction et la configuration des articulateurs lors de la production de la parole. Par exemple, /p/ est une plosive (qui nécessite une fermeture totale du conduit vocal suivi d'un relâchement) alors que /f/ est une fricative (qui nécessite une constriction sans fermeture du conduit vocal pour être produite).

[6] Un voyelle dite nasale, par opposition à une voyelle orale, est une voyelle dont la production est accompagnée du passage de l'air dans les fosses nasales (et parfois la bouche) grâce à l'abaissement du voile du palais (velum).

des phonèmes /p/ et /ɛ̃/. Le remplacement d'un phonème par un autre entraînera nécessairement la création d'un nouveau mot qui appartient ou non au lexique (e.g., /b/ et /ɛ̃/ « bain » ou /z/ et /ɛ̃/ = « zin »). C'est un apprentissage indispensable que doit faire l'enfant pour pouvoir apprendre des unités lexicales et acquérir les sens des mots nécessaires à l'établissement un vocabulaire. L'existence des phonèmes en tant qu'unités du langage et non plus en tant que sons (dans le sens psychophysique du terme) est présente très tôt dans le développement. En effet, même si les bébés ne savent pas parler, nous savons aujourd'hui qu'ils ont, à un âge très précoce, de surprenantes capacités de reconnaissance de la parole. Ils peuvent par exemple dès la naissance différencier des phonèmes d'un son non langagier (e.g., Boysson-Bardies, 1996).

1.2.1.1 DEVELOPPEMENT

1.2.1.1.1 LA PERCEPTION CATEGORIELLE

Dans les années 70, il était déjà acquis que les bébés âgés de seulement un mois étaient capable de discriminer certaines syllabes qui ne diffèrent que sur une caractéristique phonétique (e.g., le *Voice Onset Time* ou VOT[7]). Eimas, Siqueland, Jusczyk et Vigorito (1971) ont mesuré le taux de succion des bébés, celui-ci étant modulé en fonction de leur perception, indiquant qu'ils étaient capables de discriminer deux plosives (e.g., /b/ et /p/) insérées dans une syllabe.

Cette expérience a également permis de constater que les bébés ont une perception en « tout ou rien » ou « catégorielle » des sons de langage. En effet, si sur le plan physique, un /p/ et un /b/ peuvent être représentés sur un *continuum* de VOT, ils sont en revanche, sur le plan psycholinguistique, *séparés* par une barrière ou frontière catégorielle (Figure 1).

[7]Le VOT est la différence temporelle (en ms) entre l'explosion et le début de la vibration des cordes vocales (voisement). Il peut être négatif (quand la phonation précède la plosion) ou positif quand le relâchement apparait avant la phonation (Hazan et al., 2005).

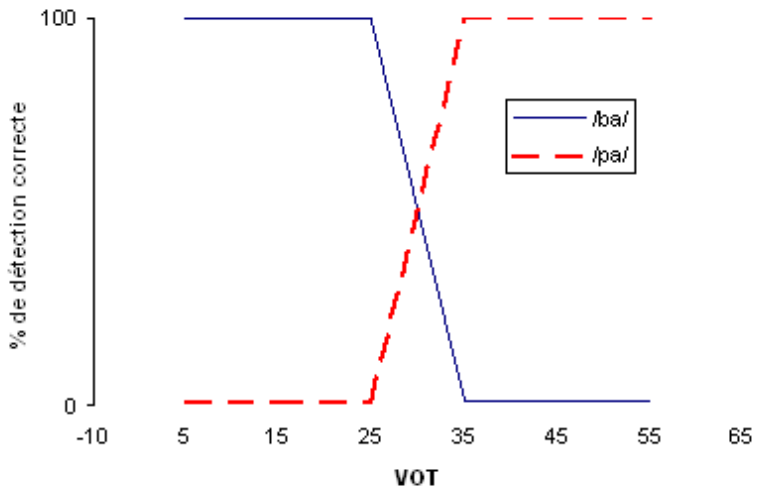

Figure1 . Pourcentage d'identification /ba/ (trait plein) et /pa/ (tirets) en fonction de la longueur du *Voice Onset Time* (VOT). (Abramson & Lisker, 1970)

Si un stimulus est dépourvu de signification phonémique, il ne sera pas perçu de manière catégorielle (Liberman, Harris, Eimas, Lisker, & Bastian, 1961). Liberman, Delattre et Cooper (1958) sont à l'origine de cette notion. En manipulant les valeurs de VOT entre un /p/ et un /b/, ils ont créé un continuum entre ces deux consonnes. Cependant ils se sont aperçus que du point de vue perceptif, la transition entre ces deux consonnes n'était jamais progressive et continue. On ne peut pas, par exemple, percevoir de phonème ou de fusion entre le /b/ et le /p/ car il y a une « frontière perceptive », arbitraire, entre ces deux phonèmes. Ces frontières sont spécifiques à chaque langue, comme l'on montré Abramson & Lisker (1970, 1973) qui ont observé les variations qui existaient entre différentes langues dans l'utilisation du VOT pour catégoriser /p/ et /b/ (Figure 2).

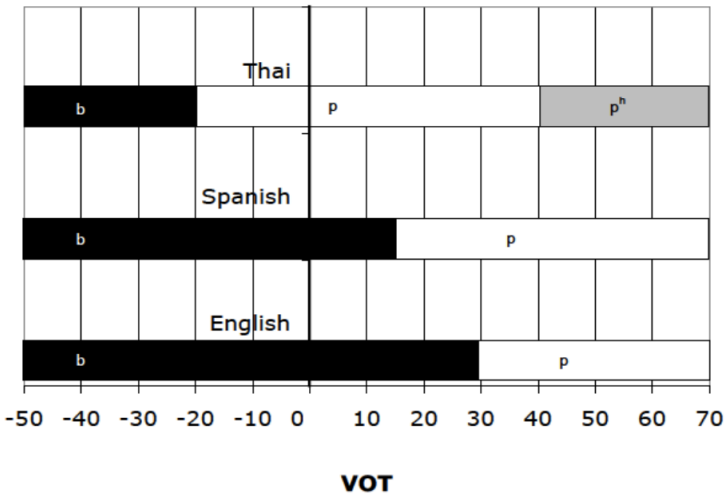

Figure 2. Différences de frontières catégorielles basées sur le VOT en fonction de la langue des individus. (Tiré des résultats de Abramson & Lisker, 1970, 1973)

Massaro & Cohen (1983) ajoutent que la perception des sons bénéficierait d'un double encodage, continu et discret (catégoriel), ce dernier mode reflétant un codage symbolique. Cela permettrait d'avoir une perception catégorielle tout en préservant les différences intra-catégorielles (Pisoni, 1990). Cela a été largement étudié par Joanne Miller en 1994 qui postule l'existence de structures internes à chaque catégorie phonémique, celles-ci étant graduées de l'exemplaire prototypique jusqu'aux exemplaires se rapprochant des frontières inter-catégorielles.

1.2.1.1.2 VERS UNE PERCEPTION SPECIFIQUE DE LA LANGUE MATERNELLE

La mise en place des catégories phonétiques se ferait sur la base distributionnelle des propriétés acoustiques de la langue maternelle : *distribution-based hypothesis* (Guenther & Gjaja, 1996 ; Jusczyk, Luce, & Charles-Luce, 1994 ; Maye, Weiss, & Aslin, 2008 ; Maye, Werker, & Gerken, 2002 ; Saffran, Aslin, & Newport, 1996). En effet la plupart des théories décrivant la mise en place des catégories de la langue maternelle (L1) (pour ne citer que quelques exemples *: Native Language Magnet* (Iverson & Kuhl, 1995) ; *Native Language Neural Commitment* (Kuhl, 2004 ; Zhang, Kuhl, Imada, Kotani, & Tohkura, 2005) ; *Linguistic Perception Model* (Escudero, 2005)) postulent que l'enfant, en étant exposé à la langue maternelle, établit des catégories correspondant à chaque son, soit par filtrage des informations non-pertinentes, soit par une stratégie qui vise à minimiser la probabilité de confusion en s'appuyant sur les indices pertinents d'une langue donnée. Cela mène à la création de « cartes perceptives » qui déterminent des frontières spécifiques entre les phonèmes pour chaque langue. Par exemple, Escudero et Polka (2003), ont montré que les locuteurs canadiens francophones produisent et perçoivent /ɛ/ et /æ/ en insistant principalement sur les différences dans les valeurs du premier formant[8] (F1) alors que les locuteurs canadiens anglophones se basent à la fois sur des indices de durée et sur la F1. Les dimensions acoustiques sont donc combinées de façon « *language-specific* » (spécifique au langage dans la suite du manuscrit).

[8] Les formants sont les zones de résonnance acoustique dans le conduit vocal matérialisé par les bandes sombre sur un spectrogramme. Le premier formant ou F1 est la première zone de résonance observable.

1.2.1.1.3 LA NOTION DE CONTRASTE PHONOLOGIQUE

Les frontières catégorielles, ainsi que la prototypicalité des exemplaires sont donc spécifiques à chaque langue. Il semblerait que ces catégories soient présentes assez précocement dans le développement, et soient influencées par l'exposition à la langue maternelle durant les premiers mois de la vie, avec des catégories langage-dépendantes qui émergent dès six mois pour les voyelles (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992) et à partir de deux mois pour les consonnes (Eimas et al., 1971). L'utilisation de certaines variations acoustiques et frontières au détriment d'autres fait référence à une notion fondamentale de la perception de la parole. Celle de co*ntraste phonologique*. Un contraste est défini par le fait qu'une substitution de phonème dans un mot produit un changement de sens. Par exemple, en français, « peau » (/po/) et « beau » (/bo/), qui se distinguent sur le contraste de voisement (i.e., le premier est non voisé alors que le second est voisé), sont perçus comme deux mots différents car /p/ et /b/ appartiennent à deux catégories phonologiques distinctes. Cela ne sera pas le cas en arabe, où /p/ n'a pas de réalité phonologique, ce qui rend donc la distinction de ces deux sons difficile (Flege & Port, 1981).

Les traits qui ne seront pas utilisés dans une distinction phonologique entraîneront parfois une perception allophonique. Un *allophone* est une des réalisations possibles d'un phonème sans toutefois que cette différence soit contrastive (différence phonétique portant sur un trait sans engendrer de contraste phonologique). Ainsi, un anglophone percevra /p^h/ et /p/ comme le même phonème /p/, le premier étant prononcé à l'attaque d'un mot (Whitley, 2002). Les variations phonétiques induites par le trait aspiré ne sont donc pas exploitées en anglais pour distinguer des phonèmes. La perception allophonique réduit donc les différences qui peuvent être perçues entre des phonèmes (Pegg & Werker, 1997). L'importance du rôle des allophones dans la perception a été montrée notamment par Boomershine, Hall, Hume et Johnson (2008), Harnsberger (2001) et Hume et Johnson (2003). Cependant, ces oppositions de traits qui constituent des allophones dans certaines langues peuvent être constitutives de phonèmes dans une autre. Par exemple, si /p^h/ et /p/ sont des allophones en anglais, ils constituent deux phonèmes différents en Hindi ou en Thai (Burnham, 2003 ; Erdener & Burnham, 2013).

D'autres caractéristiques contrastives sont également mises en place précocement. Par exemple, entre six et neuf mois, les nourrissons deviennent sensibles à la place de l'accent tonique dans les mots de leur langue (Skoruppa et al., 2009), ainsi qu'à d'autres propriétés phonologiques des mots, telles que les règles phonotactiques de la langue (Jusczyk, Cutler, &

Redanz, 1993). Avant la fin de leur première année de vie, les nourrissons ont donc acquis en grande partie le système sonore de leur langue maternelle.

Après avoir décrit les notions de base de la perception de la parole ainsi que les étapes du développement précoce de cette dernière, nous allons présenter dans les grandes lignes les modèles de perception de la parole. La question sous-jacente est de savoir quelle est la nature des unités de parole, c'est-à-dire quel type d'information est extrait et traité afin de comprendre ce que dit notre interlocuteur.

1.2.2 MODELES DE PERCEPTION DE LA PAROLE

Même si de nombreuses questions restent ouvertes quant aux mécanismes de perception de la parole, et ce, à bien des niveaux, trois écoles de pensée majeures s'affrontent encore pour déterminer la nature même de cette perception. Nous commencerons par présenter les théories auditives (Crowder & Morton, 1969 ; Pisoni, 1973 ; Stevens, 1975) et motrices de la perception de la parole (Liberman & Mattingly, 1985 ; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) et terminerons avec les théories sensori-motrices de la perception de la parole (Schwartz, Abry, Boë, & Cathiard, 2002 ; Schwartz, Basirat, Ménard, & Sato, 2012).

1.2.2.1 LES THEORIES MOTRICES ET AUDITIVES

Comme l'écrivaient déjà Liberman, Delattre et Cooper en 1952, « *we should expect that the relation between perception and articulation will be considerably simpler than the relation between perception and acoustic stimulus*[9] » (pp. 72). Cette phrase résume à elle seule le débat qui existe entre les différentes théories de la perception de la parole. L'enjeu du débat est de résoudre le problème du manque d'invariance dans la parole afin de comprendre la façon dont les innombrables réalisations d'un phonème peuvent être interprétées comme étant un seul et même phonème par l'interlocuteur. Ce phénomène est aussi connu sous le nom de « *many-to-one mapping* » qui est définit par le fait qu'un phonème (e.g., /a/) peut être réalisé de manière différente en fonction du locuteur, du débit, des phonèmes adjacents, etc (e.g., a_1, a_2, a_3, ... , a_n) tout en étant bien perçu et identifié comme un /a/ par l'auditeur. Ce manque d'invariance trouve

[9] Trad. « Nous devrions espérer que la relation entre la perception et l'articulation sera considérablement plus simple que la relation entre la perception et le stimulus acoustique ».

sa cause à plusieurs niveaux. Le premier est que la production de la parole nécessite de nombreuses étapes qui s'échelonnent de l'intention de communiquer jusqu'au recrutement des commandes motrices pour exécuter les gestes articulatoires et produire un signal acoustique. On peut comprendre « qu'entre la première étape, conceptuelle, et la dernière, articulatoire, un ensemble de paramètres linguistiques et non-linguistiques peuvent affecter la sortie » (Meunier, 2005, p. 351). Le second phénomène qui génère de la variation dans les réalisations phonétiques est la coarticulation. Celle-ci engendre des modifications dans la réalisation d'un phonème en fonction du contexte phonétique environnant. Ainsi, le phonème /d/ ne sera pas réalisé de la même manière lorsqu'il est suivi d'un /a/ ou d'un /u/ (Liberman et al., 1967).

Face à ce constat, deux principaux courants de pensée s'opposent sur l'essence même des informations utilisées pour comprendre la parole : auditive ou articulatoire. Les théories auditives de la perception de la parole considèrent que l'objet de la perception de la parole est auditif : l'invariant serait dans le signal acoustique (Crowder & Morton, 1969 ; Pisoni, 1973 ; Stevens, 1975). L'auditeur retrouverait le message en accédant, à partir d'un traitement auditif de base, commun aux mammifères, aux indices du signal qui permettent l'activation mentale de labels phonétiques ou prototypes. Ces théories ne tiennent donc pas compte du versant articulatoire de la parole. A l'opposé, la théorie motrice de la perception de la parole, développée dans les années 50 aux Laboratoires Haskins, et portée entre autres par Liberman et collègues (Liberman & Mattingly, 1985 ; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) postule qu'un auditeur se base sur sa propre expérience de production pour comprendre et identifier un phonème produit par le locuteur. L'invariant se situe donc au niveau des commandes neuro-musculaires transmises aux articulateurs et permet de « désambiguïser » le signal acoustique. Un module spécifique serait donc dédié à l'analyse de la gestuelle réalisée par le locuteur. Notons que dans cette version de la théorie motrice, ce sont les gestes intentionnels du locuteur (« *intended gesture* ») qui sont codés plutôt que les mouvements réels des articulateurs. La théorie motrice de la perception de la parole suppose l'existence d'un système de perception inné et spécifique au langage.

Cette théorie est à rapprocher de celle formulée par Fowler en 1986, la « Théorie Réaliste de la perception Directe de la parole » (« *Direct Realist Theory of speech perception* », Fowler, 1996, 1986 ; Galantucci, Fowler, & Turvey, 2006). Celle-ci suppose un accès direct du monde par les sens. Cependant, elle s'oppose à la précédente sur deux points fondamentaux. Fowler propose que ce ne serait pas la récupération des commandes neuro-musculaires liées aux gestes de parole qui permettrait la perception de la parole, mais plutôt une information

directe, tirée de la structure du conduit vocal (ou autre médium dans le cas de sons non-langagiers), c'est-à-dire les mouvements véritables du conduit vocal qui seraient codés. De plus, cette théorie postule qu'il n'y aurait pas de module spécialisé (et s'oppose donc sur ce point à la théorie motrice de Liberman et collègues). Ainsi, les mêmes mécanismes seraient à l'œuvre pour la perception d'autres modalités. Ces deux théories admettent cependant que le système moteur est sollicité lors de la perception de la parole.

Certains résultats contradictoires aux postulats des théories motrices ainsi que la complexité des modèles proposés atténuèrent pendant quelques années l'intérêt de ces théories. En effet, des études montrant que les cailles ou les chinchillas étaient capables de perception catégorielle révélaient en effet que la composante gestuelle ne pouvait pas expliquer la perception catégorielle chez l'animal (Kluender, Diehl, & Killen, 1987 ; Kuhl & Miller, 1975, 1978). Il est cependant bon de signaler qu'alors que les bébés humains n'ont besoin que de quelques minutes pour faire preuve de perception catégorielle lors de la présentation de stimuli langagiers, les animaux ont besoin de plusieurs milliers d'essais avant d'obtenir les mêmes performances. D'autres résultats expérimentaux mirent également les théories motrices à rude épreuve. En effet, en 1980, Mann montrait une adaptation des auditeurs aux modifications de réalisations acoustiques en fonction du contexte dans la production des phonèmes, ce qui était alors considéré comme une preuve de l'utilisation inconsciente de l'articulation du locuteur. Cependant, cette même adaptation a par la suite été observée dans des contextes non-langagiers (Lotto & Kluender, 1998).

A la fin des années 90, les théories motrices jouissent d'un regain de popularité suite à la découverte des « neurones miroirs » (Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). Ces auteurs ont mis en évidence que des neurones situés dans le cortex pré-moteur du macaque répondaient à la fois lorsque le singe effectuait une action spécifique (e.g., saisir un objet) mais également lorsqu'il observait la même action réalisée par l'expérimentateur. Cette découverte, au départ fortuite, a fourni une preuve neuro-anatomique quant à l'implication de certaines zones motrices aussi bien pour la réalisation d'une action que lors de la perception de celle-ci. Par la suite, des études ont également mis en évidence l'implication du système moteur lors de la perception du langage oral (Fadiga, Craighero, Buccino, & Rizzolatti, 2002 ; Pulvermüller et al., 2006 ; Wilson, Saygin, Sereno, & Iacoboni, 2004). Même si l'existence de « neurones miroirs » chez l'homme est toujours sujette à débat dans la littérature (Hickok, 2009 ; Rizzolatti & Craighero, 2004) cette hypothèse va en faveur de l'implication des représentations motrices dans la perception de la parole (Galantucci et al., 2006). Enfin, des études montrant

qu'il est possible d'utiliser différentes stratégies articulatoires pour réaliser un même phonème (Mielke, Baker, & Archangeli, 2010 ; Savariaux, Perrier, Orliaguet, & Schwartz, 1999), allant plutôt dans le sens des théories auditives, et des études montrant l'activation des cortex moteur et prémoteur lors de la perception de la parole (allant plutôt dans le sens des théories motrices) ont permis de conclure que ni les théories motrices ni les théories auditives de la perception de la parole ne peuvent expliquer seules l'ensemble des mécanismes impliqués dans la perception de la parole.

1.2.2.2 UNE APPROCHE SENSORI MOTRICE DE LA PERCEPTION DE LA PAROLE

La théorie de la perception pour le contrôle de l'action a alors été proposée par Schwartz, Abry, Boë et Cathiard (2002) et Schwartz, Basirat, Ménard et Sato (2012). Celle-ci propose une vision sensorimotrice de la perception de la parole, donnant de l'importance aux entrées auditives mais ajoutant que ces dernières ne peuvent être décodées qu'à partir des connaissances motrices procédurales que partagent le locuteur et l'auditeur. En d'autres termes, les représentations articulatoires de l'auditeur contraindraient partiellement l'interprétation des signaux acoustiques. Dans cette théorie, la parole est considérée tout aussi bien sur le versant de la production que de la perception. Deux concepts sont ici centraux : « *Perception shapes action* » (la perception façonne l'action), qui fait référence au fait que la production du langage serait influencée par la perception de celle-ci, et de manière réciproque « *Action shapes perception* » (l'action façonne la perception), où la perception de la parole est contrainte par l'action. Pour illustrer le premier concept, les auteurs se basent sur la transition vocalique /i/ → /y/, qui génère un percept discret lorsqu'elle est prononcée alors que la transition articulatoire et continue. Plusieurs configurations motrices différentes correspondent donc au même percept. Ce ne sont donc pas uniquement les représentations motrices qui guident la perception. Le second principe trouve un argument dans des données neuro-anatomiques (i.e., les neurones miroirs cités précédemment) et comportementales (Galantucci, Fowler, & Turvey, 2006 ; Rosenblum, Miller, & Sanchez, 2007 ; Rosenblum, 2005, 2008). Elles suggèrent que des connections entre les systèmes de perception et de production existent et que celles-ci activeraient les représentations motrices lors de la perception. Cela peut également se placer dans le cadre de la perception des phonèmes qui n'existent pas dans notre langue, où, dans ce cas, des connaissances motrices procédurales ne sont pas partagées par les interlocuteurs. En effet, lors de la perception de phonèmes qui n'existent pas dans notre

répertoire phonologique nous sommes contraints à identifier des phonèmes que nous ne savons pas produire.

De nombreux modèles de perception de la parole s'accordent néanmoins sur le fait que le système de perception de la parole se focalise, au cours du développement, sur les phonèmes de la langue avec laquelle l'enfant est en contact. Sur ce point, Eimas (1975), avec la notion de détecteurs de caractéristiques phonétiques, rejoint Liberman et al. (1967) et la théorie motrice. Le développement de la parole serait basé sur une sélection/maintien des caractéristiques natives ; cela va de pair avec une réorganisation perceptive qui nous fait filtrer les caractéristiques acoustiques non natives. Mais alors en quoi devenir expert dans la perception de sa langue maternelle nous empêche d'appréhender correctement une langue étrangère ? Cette question est l'objet de la section suivante, qui passe en revue les mécanismes qui sont mis en œuvre lors de la perception des phonèmes d'une langue étrangère ainsi que les modèles qui tentent de comprendre les origines de ces difficultés.

1.3 LA PERCEPTION DES PHONEMES NON-NATIFS

« Those who know nothing of foreign languages
know nothing of their own. »
— Johann Wolfgang von Goethe

1.3.1 GENERALITES: CE QUE NOTRE CERVEAU ENTEND

Durant notre développement, l'expérience que nous avons de notre langue maternelle détermine la création de catégories phonologiques spécifiques à notre langue (Werker & Tees, 1984). Pour appréhender efficacement une langue étrangère, il nous faut avant même d'acquérir un vocabulaire ou des règles syntaxiques, apprendre à percevoir des sons qui ne font pas partie de notre répertoire phonologique. L'identification des phonèmes lors de l'apprentissage d'une langue étrangère (L2) est cruciale, car si elle est mise en échec, les traitements lexicaux qui en découlent seront d'emblée biaisés. Cependant, comme le font remarquer Catherine T. Best et Mickael D. Tyler : « *Our language experience systematically constrains perception of speech contrasts that deviate phonologically or phonetically from those of the listener's native language*[10]» (2007, p.1).

En effet, la difficulté lorsque l'on perçoit une langue étrangère est principalement due au fait qu'un phonème ou contraste phonologique qui n'existe pas dans notre répertoire sera mal interprété car celui-ci dévie des représentations phonologiques que nous avons construites jusqu'alors. Cependant, toutes les langues, même si elles partagent généralement bon nombre de phonèmes, ont également un répertoire phonologique propre. Elles diffèrent dans le nombre et la nature des sons qui sont utilisés pour induire un sens et cela modifie la façon dont les auditeurs perçoivent ces sons (Bradlow, 1993 ; Fox, Flege, & Munro, 1995 ; Levy & Strange, 2008 ; Polka, 1995 ; Scholes, 1967, 1968 ; Terbeek, 1977). De ce fait, comprendre une conversation dans notre langue maternelle est aussi élémentaire qu'il est coûteux de comprendre un étranger pour une oreille, ou plutôt pour un cerveau non exercé.

[10] Trad. « Notre expérience linguistique contraint systématiquement la perception des contrastes qui dévient phonologiquement ou phonétiquement de ceux de la langue maternelle de l'auditeur ».

1.3.1.1 LA SURDITE PHONOLOGIQUE

Notre expérience de la langue maternelle modèle la façon dont notre système perceptif interprète les informations acoustiques qu'il considère comme pertinentes dans le cadre langagier. Cette spécialisation n'a pas de répercussions seulement à un niveau langagier mais existe pour beaucoup de mécanismes perceptifs. Les enfants seraient au départ des êtres sensoriellement « aspécialisés » et pourraient par conséquent discriminer tous les visages (Pascalis, Haan, & Nelson, 2002), rythmes musicaux (Hannon & Trehub, 2005) et autres contrastes qui existent dans le monde, ce que les adultes non sensibilisés ont beaucoup plus de mal à faire. En effet, en étant exposé aux configurations sonores, visuelles, etc, de son environnement, l'enfant se spécialise dans chacun de ces domaines (e.g., la perception des phonèmes de langue maternelle) et devient « insensible » à certains autres types de différences, telles que les variations phonétiques qui existent dans les autres langues. Il se « focalise » sur les informations pertinentes de la langue à laquelle il est exposé (Jusczyk & Luce, 2002 ; Kuhl et al., 2008), restreignant par là même sa perception des phonèmes qu'il na pas l'habitude d'entendre. C'est ce que l'on nomme *l'affinage perceptif*. Cependant, et comme le disent Pons et al. (2009), « *It is important to emphasize that perceptual narrowing does not reflect a complete loss of perceptual sensitivity to non-native sensory inputs; rather, it reflects a reorganization of perceptual mechanisms that then leads to decreased sensitivity to non-native sensory inputs*[11]» (p. 10598). Ce réalignement (Abramson & Lisker, 1970) des capacités perceptives dans le cadre des langues étrangères mène à la surdité phonologique (SP) (Polivanov, 1931). Ce phénomène s'installe de manière progressive durant le développement et a notamment été décrit par Troubetzkoy qui fût le premier à la considérer dans son *Grundzüge der Phonologie* en 1939. Il utilise la notion imagée de crible (phonologique) pour la représenter. En d'autres termes, cela implique que les phonèmes qui n'existent pas dans notre langue sont perçus au travers du filtre des catégories phonologiques natives. Cela entraîne une grande difficulté, voire une incapacité à distinguer des phonèmes qui ne sont pas présents dans la langue maternelle (L1). Elle touche même des phonèmes communs aux deux langues mais dont la réalisation phonétique diffère.

Ce sont Werker et Tees, qui ont mené en 1984 la première étude ayant permis de mettre expérimentalement en lumière deux éléments développementaux fondamentaux de la perception de la parole chez le bébé. Leur étude avait pour but de tester les capacités de

[11] Trad. « Il est important d'insister sur le fait que l'affinage perceptif ne reflète pas une perte complète de la sensibilité aux entrées sensorielles non natives ; elle refléterait plutôt une réorganisation des mécanismes perceptifs qui entraîne une diminution de la sensibilité à des entrées sensorielles non natives ».

discrimination de consonnes natives et non natives de bébés de 6, 10 et 12 mois. Les contrastes étudiés étaient composés de consonnes glottiques vélaire /k'/et uvulaire /q'/ contrastives en langue salish, et des consonnes non voisées rétroflexe /ʈ/et dentale /t/ de l'hindi.

A six mois, les enfants anglophones étaient capables de faire la différence entre les deux consonnes de l'hindi et celles du salish qu'ils n'avaient jamais entendu jusqu'alors (Figure 3). A 10 mois cependant, ces capacités commençaient à diminuer significativement jusqu'à disparaitre à 12 mois. A cet âge, alors que les enfants salish ou hindi conservaient la capacité à discriminer les phonèmes consonantiques présents dans leur langue, les enfants anglophones en étaient désormais incapables.

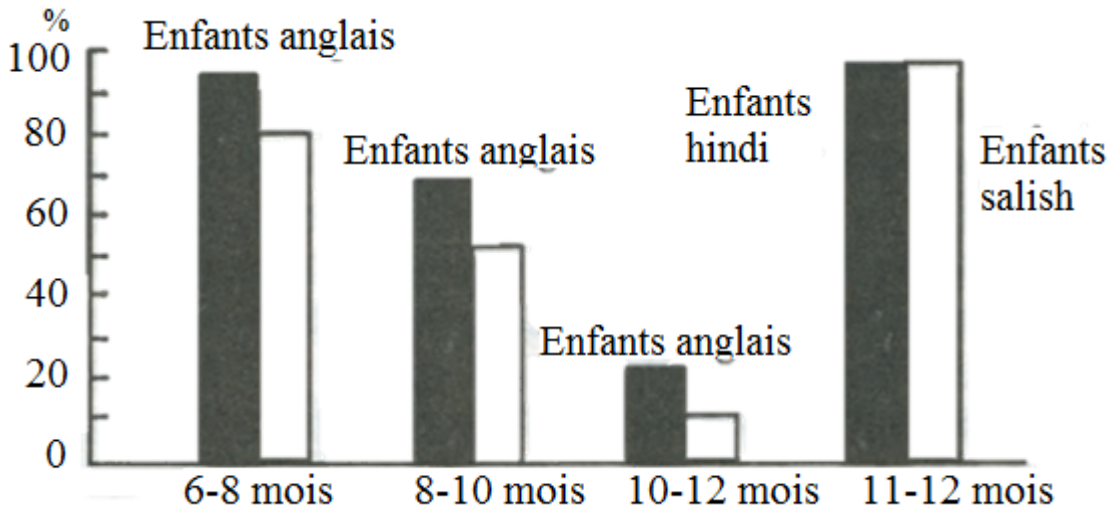

Figure 1. Pourcentage de réponses correctes en fonction des trois groupes d'âge et de la langue. Les barres noires représentent le contraste hindi ; les barres blanches représentent le contraste salish. (Extrait de Werker & Tees, 1984).

Dans la même veine, Kuhl et al., ont, en 2006, montré que les enfants japonais et anglais de 6-8 mois étaient capables de discriminer le contraste anglais /ra-la/ qui n'existe pas en japonais. Les capacités de discrimination persistaient à 10-12 mois pour les anglophones alors que les bébés japonais du même âge étaient devenus incapables de distinguer les deux phonèmes qui n'existent pas dans leur répertoire phonologique. Ces enfants étaient devenus « sourds » aux contrastes qui n'ont pas de réalité phonologique dans leur langue maternelle. Au début de leur développement, les bébés sont donc capables de discriminer tous les contrastes phonologiques qui existent dans les langues du monde. A la fin de la première année de vie, seuls les contrastes ayant une « utilité » et ayant été assez rencontrés dans l'environnement seront maintenus, les autres seront « oubliés » (Polka & Werker, 1994 ; Werker & Tees, 1984). Cet impact de la langue maternelle sur la catégorisation phonologique affecte donc la perception des contrastes dans le sens d'un maintien des contrastes natifs et d'un déclin des contrastes non-natifs (voir Saffran, Werker & Werner, 2006 pour une revue). Il

serait observable dès six mois pour la perception des voyelles (Polka & Werker, 1994 ; Kuhl, et al.,1992) et dès 10-12 mois pour les consonnes (Kuhl et al., 2006). Certaines de ces études montrent également une progression de la sensibilité menant à une facilitation de la discrimination de contrastes natifs (Kuhl et al., 2006 ; Narayan, Werker, & Beddor, 2010).

Il est à noter que cette « surdité » touche tous les attributs des phonèmes, incluant également la place d'accentuation ainsi que la longueur des voyelles et consonnes (qui sont distinctives en anglais et en japonais notamment). Par exemple, Dupoux et Sebastián-Gallés (1997) ont montré que si l'accentuation est utilisée dans la L1, elle sera plus susceptible d'être exploitée pour différencier des contrastes qui varient sur cet attribut. Si l'accentuation n'a pas de valeur phonologique, comme en français, qui est une langue à accent fixe, alors aucune différence ne sera faite à partir de cet indice (pour une expérience sur la perception des voyelles anglaises non accentuées par les Allemands, voir Braun, Lemhöfer, & Mani, 2011). De la même façon, en 1999, Dupoux, Kakehi, Hirose, Pallier et Mehler ont montré que les Français avaient beaucoup de difficultés à distinguer des séquences qui ne varient que sur la durée, contrairement aux Japonais pour lesquels cet indice a une valeur lexicale.

Quels mécanismes sous-tendent cette incapacité à exploiter des traits phonétiques ou plus largement à percevoir des phonèmes qui n'existent pas dans notre langue maternelle ?

1.3.1.2 ASSIMILATION PHONOLOGIQUE

Quand des phonèmes non-natifs partagent des traits phonétiques qui existent dans notre répertoire phonologique, nous avons tendance à les assimiler avec ceux qui existent dans notre propre répertoire, c'est-à-dire, à les considérer comme appartenant à une catégorie erronée. Prenons l'exemple du contraste /θ/-/s/ qui existe en anglais mais pas en français (la place d'articulation interdentale n'est pas exploitée en français). Lorsqu'un francophone perçoit le mot anglais « *thick* » (/θIk/, épais), il doit, pour le comprendre, identifier les phonèmes qui le composent. Comme /θ/ n'existe pas dans son répertoire, il n'a d'autre alternative que de classer ce phonème dans une catégorie préexistante de son lexique phonologique. Le /s/ sera dans ce cas, le phonème le plus similaire (il partage le mode d'articulation, car les deux sont des fricatives (non voisées), et sa place d'articulation est alvéolaire, donc « proche » de l'interdentale) auquel /θ/ pourra être assimilé. Il percevra donc « *thick* » comme « *sick* » (/sIk/, malade). Cela est en partie dû au fait que dès lors que nous percevons de la parole, nous associons automatiquement les unités phonologiques qui la composent à une des catégories qui existent dans notre répertoire phonologique. De manière générale, lorsqu'un phonème qui

n'existe pas dans le répertoire phonologique natif est perçu, celui-ci sera *assimilé* à une catégorie phonologique de la L1, même s'ils sont phonétiquement différents. Toute incertitude du système sera « compensée » par une assimilation.

Ce phénomène s'observe aussi bien pour les consonnes que pour les voyelles. Pour ces dernières, certains exemples sont assez spectaculaires. On peut par exemple citer l'étude de Lengeris, (2009), où les 11 monophtongues[12] anglaises sont assimilées aux cinq voyelles du grec moderne /i e a o u/. Pour les consonnes, des études montrent par exemple que les allemands, qui n'ont pas de contrastes /ɕ/-/s/ (contrairement aux Polonais) assimilent le /ɕ/ à leur /s/ (Lipski & Mathiak, 2007). Iverson et al. (2008) constatent la même assimilation pour le contraste /v/-/w/ anglais par les allemands et les sinhalas. Hamann & Sennema (2005) montrent quant à eux que les allemands ne considèrent pas le voisement comme ayant une valeur contrastive. En effet, même s'ils apprennent le suisse, qui a trois labiodentales (i.e., fricative non voisée /f/, fricative voisée /v/, et l'approximante voisée /ʋ/), ils assimilent la fricative labiodentale voisée à l'approximante. Il est à noter que tous les exemples cités ci-dessus ne considèrent que les cas qui impliquent la perception de phonèmes inconnus. Tout un pan de la littérature concernant la surdité phonologique et l'assimilation s'intéresse également aux difficultés subies lors de la perception de phonèmes non-natifs qui existent dans le répertoire phonologique natif mais dont la réalisation phonétique varie. En effet, les modèles décrivent généralement des scénarii différents dans les deux cas. Pour notre part, nous nous focaliserons préférentiellement sur les scénarii dits "nouveaux" dans lesquels deux sons de la L2 sont assimilés à une seule catégorie de la L1.

1.3.1.3 FACTEURS DE VARIATION

Malgré le fait que ce phénomène d'assimilation soit robuste, il est bon de garder en tête que tous les phonèmes ne sont pas logés à la même enseigne du point de vue de la discriminabilité. Pour qu'il puisse y avoir un phénomène d'assimilation, il est nécessaire que les phonèmes soient perçus comme des sons de langage. En effet, on observe des capacités de discrimination intactes lors d'une comparaison de clics en zoulou (Best, McRoberts, & Sithole, 1988) car ceux-ci sont perçus et discriminés sur une base plus acoustique que phonologique, ces phonèmes n'étant pas considérés par un auditeur étranger à ces langues comme des sons de langage (Best & Avery, 1999). Certains contrastes sont également plus faciles à percevoir que

[12] Par opposition à la diphtongue, une monophtongue est une voyelle « unique », constituée d'un seul élément vocalique. Son articulation commence et se finit au même endroit. Ici /i: ɪ e a: ʌ ɜ: ɑ ɔ: ʊ u: æ/.

d'autres (Best, McRoberts, & Goodell, 2001; Best, 1991; Strange, 1986). Best et al., (1990) citent par exemple le contraste éthiopien d'éjective qui n'existe pas en anglais et qui est pourtant bien perçu par les enfants et les adultes. De la même façon, Polka et Bohn (1996) n'observent ni d'effet d'âge ni de langue sur des allemands et des anglais lors de la perception du contraste anglais /ɛ/-/æ/ et du contraste allemand /u/-/y/. Inversement, certains contrastes et phonèmes sont difficilement perçus. Par exemple, Aoyama et al. (2004) ont montré que le phonème /r/ était plus difficile à apprendre par des japonais que /l/. Le degré de similarité perçue influence la facilité qu'aura un apprenant à apprendre un nouveau contraste (Best, 1995). Le contraste /e/-/ɛ/ est également très difficile à acquérir par des hispanophones car /ɛ/ sera assimilé à /e/ ce qui rend leur distinction relativement difficile et ce, même pour des bilingues catalan-espagnol précoces (Navarra, Sebastián-Gallés, & Soto-Faraco, 2005).

Les difficultés de discrimination ne sont donc pas uniquement liées à la simple absence d'un phonème dans notre répertoire phonologique, d'autres facteurs jouent également un rôle. L'étude de Johnson et Babel (2010) nous éclaire sur les facteurs qui affectent la similarité perçue entre phonèmes pour les contrastes qui n'existent pas dans la langue maternelle et nous renseigne sur les paramètres qui modulent la similarité perçue entre phonèmes natifs et non-natifs. Dans cette étude, les auteurs testent un panel de six consonnes (/f θ s ʃ x h/) présentées dans trois contextes vocaliques différents (/aCa/, /iCi/ et /uCu/). Seules les fricatives non voisées /f s x h/ existent en néerlandais alors que /f θ s ʃ h/ existent en anglais américain. Dans une tâche de jugement de similarité (échelle de 1 à 5, 1 étant une similarité parfaite), ils répliquent les résultats montrant l'impact du répertoire phonologique de la L1: la similarité perçue est plus importante pour les néerlandais lors de la perception des paires /θ/-/s/ et /θ/-/ʃ/. L'effet du répertoire (ou en d'autres termes l'existence ou non d'un phonème dans le répertoire) n'explique pas toutes les variations en termes de discrimination et même si son effet a été largement documenté, il ne peut à lui seul expliquer tous les résultats. Par exemple, on observe un pattern surprenant lors de la comparaison de /x/ et /h/. En effet, les anglais qui n'ont pas le phonème /x/ obtiennent le même score de similarité que les néerlandais lorsqu'il est comparé à /h/ ou à tout autre phonème. Cela nous amène au deuxième facteur qui impacte la similarité perçue : la " distance auditive ". Cette distance a été évaluée dans une tâche AX de rapidité dans laquelle les participants devaient répondre en moins de 500 ms et dire si le stimulus X était le même que le stimulus A. Les résultats ne montrent aucune différence spécifique au langage entre les deux groupes, les variations observées n'étant donc pas dues à la langue parlée mais simplement aux caractéristiques acoustiques des stimuli. En effet, la distance

perçue à l'intérieur d'un contraste est fortement influencée par la distance acoustique brute entre les séquences. C'est pour cela que les auteurs observent une forte corrélation entre le jugement de similarité et le temps de réaction (en considérant que le temps de réaction reflète une perception non-linguistique des contrastes), appuyant le fait que ce jugement est largement basé sur la similarité *acoustique* des signaux. En effet /h/ et /x/ sont très similaires acoustiquement et sont notés comme « similaires » par les deux groupes (3/5), alors que les phonèmes /x/ et /s/ sont perçus comme différents (5/5) malgré le fait que /x/ n'existe pas chez les anglophones.

Ces résultats montrent que les différences de jugement de similarité ne sont pas dues à un traitement linguistique auditif de bas niveau. Les effets de leur étude seraient donc liés à des comparaisons acoustiques non linguistiques auxquelles s'ajoutent des effets spécifiques au langage. Best (1995) a également observé l'impact de la similarité perçue sur les capacités de discrimination des phonèmes. Johnson & Babel (2010) montrent également que des néerlandais, qui n'ont pas de distinction phonémique pour /s/ et /ʃ/ (/ʃ/ n'existe pas en tant que phonème en néerlandais mais est perçu et produit dans certain contexte comme un allophone de /s/), perçoivent ces fricatives comme plus similaires que des anglophones américains. Ils confirment donc que les *patterns allophoniques* jouent un rôle dans la perception de similarité entre phonèmes. En effet, le fait que les phonèmes /s/ et /ʃ/ présentés en contexte /i/ et /a/ soient jugés plus similaires par les néerlandais est observé car ces consonnes ne contrastent jamais dans ce contexte. A contrario, ces deux consonnes sont jugées aussi similaires par les anglophones que les néerlandophones lorsqu'elles sont présentées dans un contexte /u/. Les *patterns allophoniques* contraignent donc la similarité perçue entre des phonèmes. En effet, les phonèmes (e.g., /s/) dont les variations phonétiques ne sont pas utilisées dans une langue (e.g., /ʃ/) et qui sont donc des allophones, sont perçues comme étant davantage similaires. Boomershine et al. (2008) avaient par exemple étudié la similarité perceptive des consonnes /d/, /ɾ/, and /ð/ par des auditeurs espagnols et anglais. En anglais, /d/ et /ɾ/ sont des allophones alors que /ð/ est un phonème, et en espagnol, /ɾ/ est un phonème alors que /d/ et /ð/ sont des allophones. Les allophones ont été jugés comme étant davantage similaires entre eux dans les langues respectives qu'entre les deux langues. Cela a, comme dans l'étude de Johnson et Babel, entrainé une augmentation des temps de réponse lorsque les deux allophones devaient être jugés comme étant différents.

Les auteurs en viennent donc à la conclusion que trois composantes majeures déterminent la similarité phonétique, et donc les capacités de discrimination d'un contraste phonologique, lors de la perception de la parole : la distance auditive, le répertoire phonologique et les patterns

allophoniques. La qualité de la perception est aussi modulée par le contexte (Polka, 1991, 1992 ; Werker & Tees, 1984), les organes phonatoires impliqués lors de la production (Kuhl et al., 2006) ou encore les règles particulières aux langues en question (Kochetov, 2004).

La surdité phonologique repose donc sur le fait que la perception est fondée sur nos représentations mentales, celles-ci s'accordant, lors de la création du lexique, aux sons qui font sens dans la langue de l'enfant. Trois principaux modèles tentent de décrire les relations entre le système phonologique de la L1 et celui de la L2 (en production et en perception) et comment celles-ci vont engendrer des difficultés à percevoir correctement un contraste phonologique non-natif.

1.3.2 MODELES D'ASSIMILATION PHONOLOGIQUE

Tous les modèles récents argumentent que l'expérience de la langue maternelle façonne une « grille perceptive » qui modifie notre façon de percevoir les contrastes qui ne nous sont pas familiers. Les trois modèles qui vont être présentés dans cette section décrivent le phénomène d'assimilation. Sans être contradictoires, chacun s'intéresse à une facette particulière de ce phénomène. Le « *Perceptual Assimilation Model* » (PAM) de Best (Best, McRoberts, & Sithole, 1988 ; Best, 1994, 1995) se focalise sur la perception des phonèmes à l'intérieur d'un contraste par des individus inexpérimentés, alors que le « *Speech Learning Model* » (SLM) de Flege (Flege, Munro, & MacKay, 1995 ; Flege, 1995; Guion, Flege, Akahane-Yamada, & Pruitt, 2000) s'intéresse aux difficultés de production et de perception de phonèmes isolés par les apprenants avancés. Enfin, le « *Native Language Magnet* » (NLM) de Kuhl (Grieser & Kuhl, 1989 ; Iverson & Kuhl, 1996 ; Kuhl, 1991) s'interroge quant à lui sur la structuration des catégories phonologiques natives. Ces modèles, même s'ils ne sont pas les seuls à expliquer les difficultés lors de la perception ou de la production de phonèmes non natifs, (citons par exemple, le « *Ontogeny Phylogeny Model* » (Major, 2001, 2007), le « *Phonological Interference* » de Brown (1998, 2000), ou le « *Linguistic Second Language Perception* » de Escudero (Escudero & Boersma, 2004 ; Escudero, 2005), sont les plus influents du domaine. Ils postulent des comportements différents en fonction de la catégorie de l'assimilation qui sera dépendante du fait que les phonèmes (ou les traits qui les composent) existent ou non dans la langue maternelle et de leur similarité avec les phonèmes du répertoire.

1.3.2.1 MODELE D'ASSIMILATION PERCEPTIVE « PAM »

Le « *Perceptual Assimilation Model* » ou PAM (Best, McRoberts, & Goodell, 2001 ; Best & McRoberts, 2003 ; Best, McRoberts, & Sithole, 1988 ; Best & Tyler, 2007a ; Best, 1994, 1995) postule que les sons de la L2 seraient classés dans des catégories différentes en fonction de leur similarité intra et inter langue. Les sons qui sont le plus similaires de ceux de la L1 seraient les plus difficiles à apprendre. *A contrario*, plus ils sont différents et plus l'apprenant les placera dans des catégories différentes. La perception est analysée en termes de similarités/différences articulatoires entre L1 et L2. Quand un individu rencontre un nouveau contraste phonologique, cinq scenarii s'offrent à lui (Figure 4).

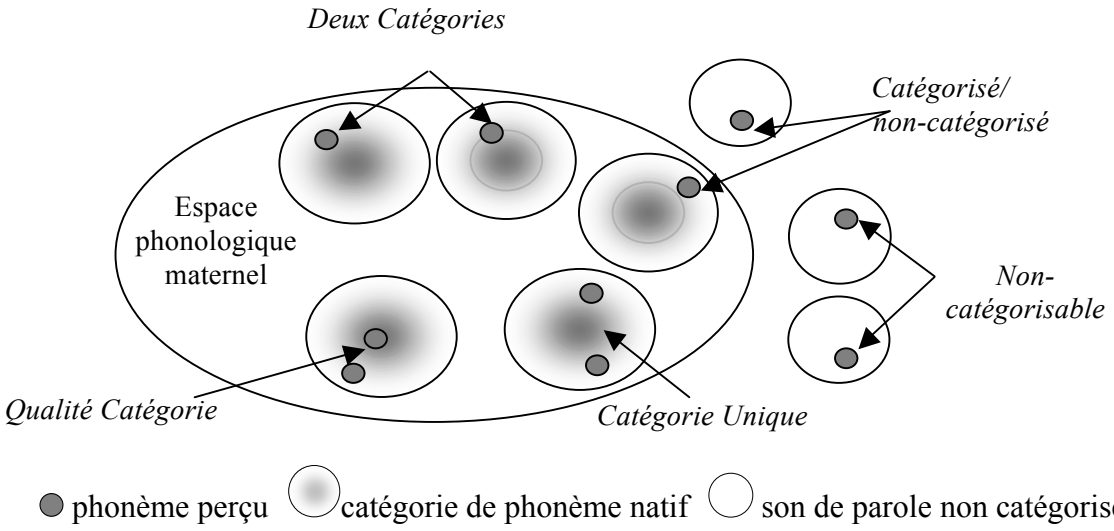

Figure 2. Représentation des scénarii prévus lors de l'assimilation de contrastes non-natifs par le modèle PAM.

Si les deux phonèmes qui constituent le contraste de la L2 sont assimilés à deux catégories natives différentes, on parle d'assimilation *« Deux Catégories »*. Ils sont donc facilement discriminables. L'exemple donné par Best (1991) est celui du contraste Hindi /ɖ/-/d_h/ qui est assimilé au contraste /d/-/ð/ anglais. Les deux phonèmes seront donc catégorisés dans deux catégories natives distinctes et même s'ils sont mal catégorisés, les anglophones parviendront à faire la différence entre ces deux phonèmes. Le deuxième cas possible est celui de l'assimilation en « Qu*alité Catégo*rie ». Il s'observe quand deux phonèmes non natifs sont assimilés à une seule catégorie native, mais qu'un des phonèmes est perçu comme proche du natif et que l'autre est jugé déviant. La discriminabilité des phonèmes est en générale assez bonne car l'un est un bon exemplaire et l'autre est déviant. C'est par exemple le cas du contraste /i/-/I/ de l'anglais. Lorsqu'ils sont perçus par des espagnols, les deux phonèmes sont

catégorisés comme un /i/ espagnol, mais /I/ est perçu comme plus déviant du prototype espagnol et peut donc être discriminé (Morrison, 2006). Quand deux phonèmes sont très difficiles à différencier, c'est qu'ils sont généralement placés dans une seule catégorie native, et considérés comme aussi déviants du natif l'un que l'autre. C'est une classification en *« Catégorie Unique »*. Enfin, deux catégories sont décrites quand les qualités du signal sont si éloignées de la structure de la parole qu'ils ne peuvent être catégorisés comme du langage. Dans un cas, l'un des deux phonèmes ne peut pas être catégorisé (*« Catégorisé/non-catégorisé »*), la discrimination est donc attendue comme très bonne. Dans l'autre cas, la discrimination est également de bonne qualité car les phonèmes sont si éloignés des propriétés articulatoires natives qu'ils ne sont pas interprétés comme étant du langage, c'est le cas de la vélaire aspirée /k_h/ et de l'éjective /k'/ en zoulou, qui sont discriminées à hauteur de 89.4% par des anglophones, alors que la plosive bilabiale voisée /b/ n'est distinguée de l'implosive /ɓ/ dans seulement 65% des cas (Best et al., 1988 ; Best et al., 2001). Cela amène au scénario *« Non-catégorisable »*. Les locuteurs n'ont pas considéré les clics comme des sons linguistiques et ont bien réussi à percevoir les contrastes (Best, 1995) (pour une revue se référer à Martin et Peperkamp, 2011).

La perception des sons étrangers serait difficile d'une part car il s'agit de faire la différence entre deux sons de la L2 et d'autre part de les différencier des sons de la L1. PAM prend en compte à la fois le niveau phonologique et phonétique. Les individus exposés à une langue étrangère ne connaissent pas encore les distinctions phonétiques qui entraînent une distinction phonologique. Un locuteur sait interpréter les variations phonétiques qui, dans sa langue, sont pertinentes au niveau phonologique, mais ce n'est pas le cas pour les sons de la L2. Il les interprète donc en fonction de ce qui fait sens pour lui. En d'autres termes, il retranscrirait les incohérences (i.e., les phonèmes qu'il ne maîtrise pas) dans ses propres catégories phonologiques. Cela lui donne un support pour appréhender un système phonologique qu'il ne connaît pas.

Dans une version améliorée de son modèle, PAM/AO (« *articulatory organ* »), Best (Best & McRoberts, 2003) suggère que le déclin des capacités de discrimination est présent quand le contraste implique des articulateurs communs, comme dans le contraste /s/-/z/ qui repose sur le voisement alors que la discriminabilité est conservée dans le cas d'un place d'articulation différent (i.e., /b/-/t/). Dans la seconde partie de leur étude, les auteurs testaient deux postulats. Alors que les enfants de 6-8 mois devaient tous parvenir à discriminer les phonèmes, les enfants anglophones de 10-12 mois devraient se comporter différemment en fonction des articulateurs impliqués dans la réalisation du contraste. Ils devraient observer une

absence de difficultés lorsque le contraste non-natif testé (i.e., bilabiale et alvéolaire éjective non voisés /pʼ/-/tʼ/) impliquait différents organes articulatoires (« *between-organ contrast* ») et une difficulté lorsque les mêmes contraintes articulatoires étaient imposées à un contraste non natif (i.e., latérale fricatives voisées et non voisées /ɬ/-/ɮ/) ou même natif (i.e., /s/-/z/) (« *within-organ contrast* »). Leurs résultats montrent une différence entre la phase d'habituation et le test pour les phonèmes qui ne partagent pas les mêmes organes articulatoires (« *between-organ contrast* ») pour les deux groupes d'âge. Concernant les *within-organ contrast* , ils sont discriminés, mais les performances sont cependant plus faibles à 10-12 mois qu'à 6-8 mois, ce qui atteste d'un déclin des capacités de discrimination pour ce type de contraste, confirmant donc les apports de PAM/AO. Il y a donc un effet non négligeable de l'impact des articulateurs pour la catégorisation des sons non-natifs, mais également natifs. Ces résultats ont été répliqués sur d'autres consonnes (Kuhl et al., 2006).

Le second modèle, présenté dans la section suivante, est moins basé sur la confrontation des contrastes de la L2 que sur la structure du système phonologique de la L1. Pour Kuhl, l'assimilation s'explique par la nature « attractive » des phonèmes de la L1, celle-ci étant dépendante de la création des catégories. Plus qu'un affinage, ce serait une hypersensibilisation aux spécificités acoustico-phonétiques du système phonologique de la langue maternelle qui entraînerait des difficultés à percevoir correctement les phonèmes de la L2.

1.3.2.2 MODELE DES AIMANTS DE LA LANGUE MATERNELLE « NLM »

Le modèle « *Native Language Magnets* » (NLM ; Kuhl, Tsuzaki, Tohkura, & Meltzoff, 1994; Kuhl et al., 1992), qui s'apparente aux « *perceptual anchors* » de la théorie de Macmillan, Goldberg et Braida (1988), met l'accent sur la structure des catégories. Le développement des catégories phonologiques maternelles se déroulerait en trois étapes et serait, entre autres, responsable des difficultés observées lors de la perception de phonèmes non-natifs. Dans un premier temps, les bébés seraient capables de discriminer tous les sons de langage, mais cette capacité serait dérivée de traitements auditifs généraux et non pas spécifiques au langage (Kuhl, 1991). Au cours de la première année de vie, avant l'acquisition de la phonologie contrastive, le bébé commence à établir des « cartes perceptives » de la parole. A force de rencontrer les phonèmes de son environnement, l'enfant construirait des représentations phonétiques basées sur des propriétés distributionnelles, donnant un poids important aux

indices pertinents de la langue maternelle. Avec l'expérience, un « modèle » pour chaque catégorie phonologique sera ainsi généré. Celui-ci jouera le rôle d'un « aimant » lors de la perception des unités du langage, les rendant, de par son action attractrice, plus similaire au modèle/prototype quel que soit leur réalisation. La sensibilité autour des prototypes sera donc moindre, favorisant une perception allophonique de par la minimisation des différences intra-catégorielles. Cette perception distendue entraînerait une facilitation lors de la discrimination des phonèmes natifs et des difficultés lors de la perception de contrastes non-natifs. La discrimination sera difficile autour des prototypes (« *native magnets* ») de la langue maternelle et l'assimilation serait due au rôle d'aimant qu'ils jouent, modifiant ainsi la perception des phonèmes étrangers dans le sens de ceux de la langue maternelle. La meilleure réalisation correspond à une réduction de la discrimination (allophone) et un regroupement perceptif.

La cause de ce phénomène serait, selon Kuhl et al. (2008), liée à des contraintes attentionnelles. Ils postulent que « *early exposure to language shapes attentional networks, and that in adulthood, they make second language learning difficult*[13] » (p. 983). Avec l'expérience, nous deviendrions insensibles à certains indices que nous n'avons jamais utilisés jusqu'alors, car ils n'existent pas dans notre langue maternelle. C'est ce que montre l'étude d'Iverson et al. (2003) menée chez l'adulte qui testait la perception du contraste anglais /r/-/l/ par des locuteurs de l'anglais, du japonais et de l'allemand. Les résultats montrent que les auditeurs se basent sur différents indices pour discriminer ces consonnes. Pour les distinguer, les japonais s'appuient sur le second formant (F2). C'est en effet un indice pertinent dans leur langue, leur attention est donc naturellement tournée de manière privilégiée vers celui-ci. Or, les anglophones se basent à la fois sur le F2 et sur le F3 pour distinguer ces deux consonnes. Le fait que les locuteurs japonais ne tiennent pas compte des indices fournis par le F3 entraîne des erreurs de catégorisation de leur part. Kuhl postule que chez l'adulte, l'engagement neural des circuits relatifs à la perception de la langue maternelle sont « rigidifiés » ce qui diminue la capacité à utiliser des indices qui n'ont jamais été utilisés jusqu'alors (cf. Chapitre 2.2.3.2 « Engagement neuronal pour la langue maternelle » pour plus de détails).

Enfin, le dernier modèle s'intéresse quant à lui à la création de nouvelles catégories phonologiques avec l'expertise.

[13] Trad. « L'exposition précoce au langage façonne les réseaux attentionnels, et à l'âge adulte, ils [les réseaux attentionnels] rendent l'apprentissage d'une seconde langue difficile ».

1.3.2.3 MODELE D'APPRENTISSAGE DES LANGUES « SLM »

Le modèle de Flege (1995, 1987) tente de prendre en compte aussi bien le versant production que perception et se focalise plus particulièrement sur le phénomène de création de nouvelles catégories avec l'expertise. Il n'est en rien contradictoire avec les deux modèles présentés ci-dessus mais s'intéresse à une autre population, à savoir les bilingues. Il est également basé sur le principe de similarité mais plutôt sur sa nature intrinsèque que sur sa nature gestuelle. Il postule que quand quelqu'un apprend une L2, s'il ne peut faire la différence entre deux sons, il les assimile à ceux de la L1 (on retrouve la même étape chez Best). La L1 et la L2 seraient issues du même espace perceptif. Les sons qui ne sont pas assimilés car trop différents de ceux de la L1 sont placés dans de nouvelles catégories. Cette théorie explique d'ailleurs la difficulté d'apprentissage d'une langue étrangère avec l'âge par le fait que plus on est âgé, plus on acquière de l'expérience avec le système phonologique de la L1 ce qui rend plus difficile la création de catégories différentes de la L1 et cause plus d'interférence. Cela ouvre d'emblée la question de l'existence d'autres facteurs, comme l'âge d'acquisition, qui peuvent moduler notre faculté de percevoir et/ou d'apprendre de nouveaux phonèmes.

1.3.3 APPRENTISSAGE DE NOUVEAUX CONTRASTES PHONOLOGIQUES

« People know two languages:

their native language and gibberish »

— Maribel C. Pagan

Nous savons aujourd'hui que la surdité phonologique n'est pas nécessairement une fatalité. Bien que la surdité phonologique soit un phénomène démontré et mis en cause de façon déterminante dans la difficulté à percevoir des sons de langues non-natives, il est évident que nous sommes capables d'apprendre des langues étrangères après l'enfance. Une grande partie de la population du monde peut communiquer dans plusieurs idiomes sans avoir nécessairement grandi dans un milieu plurilingue (Altarriba & Heredia, 2008). En effet, des facteurs ainsi que des méthodes d'apprentissage sont aujourd'hui reconnus comme améliorant les capacités d'identification et de discrimination des phonèmes.

1.3.3.1 PERIODE SENSIBLE

Le phénomène de surdité phonologique n'est pas intrinsèquement lié au langage, mais plutôt à une réduction de la sensibilité perceptive globale pour les traits auxquels nous n'avons pas été exposés. Ces soubassements seraient à la fois environnementaux et corticaux (Dehaene-Lambertz, 1997 ; Kazanina, Phillips, & Idsardi, 2006), nourrissant par là même la notion de « période critique de l'acquisition du langage ». Alors que cette hypothèse a été proposée par Penfield & Roberts (1959), le premier à utiliser ce terme dans un contexte linguistique était Lenneberg en 1967. Même si le terme de « critique » a longtemps fait débat, la définition de Oyama (1978) est celle qui fait le plus consensus : « *It is a developmental phenomenon not in that it is « determined by the genes » in some rigid or direct way, but rather insofar as it reflects an intricate sequence in interactions between the developing phenotype and the environment, which is sufficiently typical of the species that it appears despite individual differences and widely varying experiences*[14] » (p. 10).

Cette période critique est aujourd'hui assez controversée et de nombreux auteurs ont des avis divergents sur l'âge à partir duquel celle-ci se met en place, sur ses causes voire même sur la réalité de son existence. Ce phénomène implique que tous les processus implicites mis en place lors de l'acquisition de la langue maternelle se rigidifient et se limitent dès l'adolescence. Pour ce qui est de l'acquisition d'une seconde langue et sans entrer dans le débat de l'existence ou de l'âge a partir duquel cette période critique se met en place, il est admis que l'âge d'acquisition impacte les performances. De nombreuses études portant sur la production des voyelles ou des consonnes montrent en effet que plus l'âge d'acquisition d'une langue est tardif, plus les productions jugées conformes à la réalisation/fluence/grammaire des natifs diminuent (Dekeyser, 2000 ; Flege et al., 1995 ; Singleton & Lengyel Zsolt, 1995) et plus un contraste a du mal à être acquis et donc discriminé (Aoyama et al., 2004). Cela serait dû à une réduction de la plasticité cérébrale au cours du développement (Johnson & Newport, 1989 ; Pulvermüller & Schumann, 1994) ou simplement aux interférences de la L1 (Flege, Yeni-Komshian, & Liu, 1999).

[14] Trad. « C'est un phénomène développemental, non pas qu'il soit "déterminé par les gènes" d'une manière rigide ou directe, mais plutôt dans la mesure où il reflète une séquence complexe dans les interactions entre le phénotype en développement et l'environnement, qui est suffisamment typique de l'espèce pour apparaître malgré les différences individuelles et les variations importantes d'expériences ».

Cependant, certaines études ne vont pas forcément dans le sens d'une relation inverse entre l'âge d'acquisition et la maîtrise d'une nouvelle langue. Snow & Hoefnagel-Höhle (1978), dont le travail était au départ inscrit dans l'infirmation de l'existence de la période critique, a tout de même mis en avant un léger avantage du groupe de 12-15 ans sur les adultes dans de nombreuses tâches (répétition, imitation, discrimination auditive, répétition de phrase, etc), alors que les enfants plus jeunes 3-10 ans ont des performances inférieures. Les enfants de 12-15 ans sont également ceux qui bénéficient de l'acquisition la plus rapide. La prononciation spontanée étant la seule tâche qui montre une amélioration linéaire avec l'âge. Flege et al. (1999), ont testé 240 coréens qui vivaient aux Etats-Unis depuis 1 à 23 ans, sur leur capacité de prononciation et de morphosyntaxe. Ils ont obtenu un effet de l'âge d'acquisition, avec un accent plus fort pour les personnes ayant acquis la langue plus tard, effet qui ne se retrouve pas pour la morphosyntaxe lorsqu'ils contrôlent les facteurs confondus, aussi bien chez les enfants, les adolescents ou les adultes. Les auteurs préfèrent expliquer leurs résultats par des changements dans la façon dont les systèmes phonologiques de la L1 et de la L2 interagissent plutôt qu'en termes de maturation. Les enfants auraient plus de facilité à apprendre de nouveaux contrastes, notamment parce que leurs catégories phonologiques sont plus flexibles que les adultes. C'est la conclusion à laquelle arrivent également Walley & Flege (1999) qui montrent que les enfants de cinq ans acceptent plus de stimuli sur un continuum constitué de voyelles natives et non-natives comme étant des phonèmes natifs. La mise en place des catégories natives, même si elle commence très précocement dans le développement, nécessite une maturation longue pour arriver à une catégorisation équivalente à celle des adultes. Hazan & Barrett (2000) ont mis en évidence que les enfants, même à 12 ans, n'ont pas encore une catégorisation équivalente à celle de l'adulte (voir aussi Nittrouer, 2004). Ces résultats sont confortés par ceux de Heeren & Schouten (2010) qui observent une modification des frontières catégorielles dans le sens de celles du finnois alors que la frontière du groupe contrôle reste inchangée pour le groupe de néerlandais de 12 ans entrainé à l'identification du contraste /t/-/t:/ finnois. Cette flexibilité des catégories permet d'acquérir plus facilement de nouveaux phonèmes.

Il semblerait qu'une exposition précoce, même passive, permette également de modifier les réponses neuronales à des stimuli inconnus. L'étude de Cheour, Shestakova, Alku, Ceponiene et Näätänen (2002) est une étude longitudinale en environnement naturel. Des enfants finnois de trois à six ans étaient suivis sur une période de quatre mois après leur inclusion dans une classe française ou en garderie où le français était parlé entre 50 et 90 % du temps. Leurs résultats montrent que l'écoute passive peut être suffisante pour modifier la

MMN[15] (ou *Mismatch Negativity*) générée à l'écoute de deux contrastes vocaliques inexistants en finnois. Cela ne donne pas un avantage perceptif d'emblée aux enfants, vu que leur performances sont en général au départ moins bonnes que celle des adultes, mais leur scores s'améliorent plus que ceux des adultes (Aoyama et al., 2004), grâce à une plus grande flexibilité des catégories. Cette flexibilité serait due à l'état non mature des circuits neuronaux engagés lors de la perception des sons de la langue.

1.3.3.2 ENGAGEMENT NEURONAL POUR LA LANGUE MATERNELLE

La cause de ces difficultés rencontrées à l'âge adulte pour acquérir de nouveaux contrastes phonologiques se trouve dans le conditionnement opéré par notre langue maternelle sur la façon dont notre cerveau interprète les sons étrangers. L'exposition répétée à une langue produit inévitablement un engagement neuronal qui affecte les apprentissages futurs. Ces effets ont été confirmés chez l'adulte (Callan, Jones, Callan, & Akahane-Yamada, 2004 ; Golestani & Zatorre, 2004 ; Koyama et al., 2003 ; Perani et al., 2003 ; Sanders, Newport, & Neville, 2002 ; Zhang et al., 2005) mais également chez l'enfant (Conboy & Mills, 2006 ; Dehaene-Lambertz & Gliga, 2004 ; Mills et al., 2004 ; Rivera-Gaxiola, Klarman, Garcia-Sierra, & Kuhl, 2005). Kuhl (2000) propose que les encodages précoces suite à l'exposition à notre langue maternelle affectent nos capacités à apprendre de nouveaux contrastes phonologiques. Ce concept, nommé « engagement neural de la langue maternelle » (« *Native Language Neural Commitment* » ou NLNC) décrit les changements neuronaux opérés lors de l'exposition à la langue maternelle et reflètent les régularités statistiques et spectrales des sons natifs. Ce réseau ainsi créé renforce la détection des patterns de haut niveau de notre langue et réduit dans le même temps la sensibilité à tous les autres patterns de phonèmes alternatifs (Kuhl, 2004). Les aptitudes des enfants à discriminer des phonèmes qui n'existent pas dans leur langue maternelle s'expliquent donc par un état plus immature des circuits qui ne sont pas encore engagés.

De nombreuses études semblent en effet montrer des interférences induites par la langue maternelle lors de l'acquisition par l'adulte d'une nouvelle langue (Flege, 1995 ; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002 ; Zhang et al., 2009, 2005).

[15] Le terme MMN ou *mismatch negativity* fait référence à une onde cérébrale déclenchée suite à la perception d'un stimuli discordant présenté parmi d'autre, quelque soit a modalité de presentation (auditive, visuelle, etc). Celle-ci apparaît 150 à 250 ms après l'apparition du stimulus deviant.

Zhang et al. (2005) ont par exemple étudié les traitements phonologiques de Japonais et d'Américains dans deux études en magnétoencéphalographie (MEG). Les participants devaient identifier les syllabes /ra/ et /la/ (contraste qui n'existe pas en japonais) créées synthétiquement par manipulation du troisième formant. L'étude comportait une partie comportementale, et une partie neurophysiologique utilisant le paradigme de oddball[16]. Les résultats comportementaux et neurophysiologiques ont montré une sensibilité moindre des japonais au contraste /ra/-/la/ par rapport aux américains. De plus, les stimuli non natifs recrutaient plus de ressources cérébrales dans les deux hémisphères et les activations générées étaient plus longues dans les aires temporale supérieure et pariétale inférieure. Les stimuli contrôle (i.e., /ba/-/wa/ et des stimuli non langagiers similaires à /ra/-/la/ mais dont la fréquence fondamentale ainsi que les quatre formants avaient été remplacés par une composante sinusoïdale de même fréquence) n'ont pas induit de différences significatives à l'exception de la durée de l'effet dans le cortex temporale supérieur. Ces résultats soulignent donc l'impact d'une exposition précoce à une langue, avec un engagement neuronal pour les propriétés acoustiques propres à la langue à laquelle nous avons été exposés. Cet engagement interfère avec les traitements nécessaires à la perception des langues non-natives, les rendant moins efficaces.

D'autres modèles (McCandliss et al., 2002 ; Vallabha & McClelland, 2007) décrivent les processus d'apprentissage comme fonctionnant de manière similaire à des apprentissages hebbiens (i.e., apprentissage associatif) non supervisés. Selon Hebb (1949) « lorsque deux neurones ou systèmes sont fréquemment activés de manière simultanée, le poids de leurs connexions synaptiques serait renforcé de sorte que l'activité d'un de ces neurones ou systèmes suscite automatiquement l'activation de l'autre » (cité par Boulenger, 2006, p.87). Le modèle de McCandliss et al. (2002) propose par exemple que la plasticité synaptique hebbienne renforcerait les catégories qui ont été établies durant l'enfance. Ces « réseaux attracteurs » proposent donc une explication sous-jacente aux aimants perceptifs de Kuhl (1991). Cette hypothèse permet de prédire à l'avance quel type d'entraînement doit être utilisé pour remédier efficacement aux difficultés d'apprentissage à l'âge adulte. Il semblerait que ces méthodes d'entraînement adaptatif basées sur les modèles d'apprentissage hebbien permettent un apprentissage plus rapide.

[16] Le paradigme « oddball » (ou stimulus discordant) consiste à présenter un stimulus déviant occasionnel (ou rare) pendant la présentation d'une série de stimuli répétés d'une catégorie.

Cependant, nous savons que le cerveau humain, même s'il est moins plastique à l'âge adulte que durant l'enfance, reste capable de flexibilité. Serait-il possible d'apprendre de nouveaux contrastes malgré les changements neuronaux induits par l'exposition à notre langue maternelle ?

1.3.3.3 APPRENTISSAGE A L'AGE ADULTE

Des nombreuses études comportementales se sont attachées à réduire le phénomène de surdité phonologique, que ce soit en laboratoire, par des entraînements catégoriels ou en condition naturelle (i.e., immersion). La difficulté pour les japonais d'acquérir le contraste /r/-/l/ est sans doute l'exemple le plus connu et par conséquent le plus documenté, aussi bien pour constater cette difficulté que pour tenter d'y remédier.

1.3.3.3.1 ENTRAINEMENT

En laboratoire, des chercheurs ont tenté d'apprendre à des japonais à distinguer le contraste /r/-/l/. C'est par exemple le cas de Lively, Pisoni, Yamada, Tohkura et Yamada (1994) qui ont testé des japonais vivant à Kyoto qui avaient très peu de notions d'anglais. Ils furent entrainés durant trois semaines en utilisant des réalisations naturelles de paires minimales contenant les phonèmes /r/ et /l/, enregistrées par plusieurs locuteurs afin d'induire de la variabilité, favorisant donc un apprentissage de nature phonologique. Cette méthode est connue sous le nom de « *High Variability Perceptual Training* » (HVPT). Une comparaison pré- post- test était réalisée, et également une mesure de généralisation des compétences à de nouveaux items/locuteurs. Une augmentation de 11 % de discrimination correcte entre le pré-test et le test a été observée, et de manière surprenante, une conservation de ces compétences lors du post-test six mois plus tard.

Des études en neurophysiologie ont révélé que cette amélioration comportementale est également observable au niveau neuronal, celle-ci se manifestant par une MMN plus ample, apparaissant plus tôt et disparaissant plus tard après un entraînement d'une semaine chez l'adulte (Kraus et al., 1995). Mais, là encore, l'amélioration des performances était de l'ordre de 10%. Certains auteurs postulent que ces difficultés pourraient venir du fait que les participants ne s'appuient pas sur les bons indices pour distinguer les phonèmes (McClelland, Fiez, & McCandliss, 2002). C'est ce qu'ont montré Iverson, Hazan et Bannister (2005), Lively

et al. (1994) et Vallabha & McClelland (2007). Avec un entraînement adapté, notamment en exagérant certains indices pour permettre la discrimination, comme dans le « *child-directed speech*[17] », Vallabha & McClelland (2007) ont amélioré les capacités de discrimination des participants. Cependant, ce type d'entraînement limite généralement le potentiel de généralisation à de nouveaux items. Certains chercheurs ont également tenté d'entraîner les participants en manipulant artificiellement les indices pertinents du signal, afin de rendre leur distinctivité plus saillante. Iverson et al. (2005) ont, dans ces conditions, obtenu une amélioration globale de 10% (moyenne des résultats pour les trois méthodes utilisées), avec une amélioration de 18% lorsque le phonème était situé en position initiale. Cette amélioration est largement supérieure à celle obtenue par les autres équipes utilisant la méthode HVPT citées auparavant. Cependant, les auteurs font eux-mêmes remarquer que les stimuli naturels dans le cadre d'une présentation incluant beaucoup de variabilité restent les plus efficaces car l'augmentation des scores dans cette condition ne diffère pas de celle observée pour les stimuli manipulés.

D'autres études ont mis en place un paradigme original de jeu vidéo, où les personnages (des aliens) sont chacun associés à un phonème différent (Lim & Holt, 2011). Ces personnages sont utilisés pour générer des associations multimodales et donner un support contextuel aux associations. En augmentant la variabilité du deuxième formant, les auteurs ont tenté de pousser les participants japonais à se focaliser sur la F3, qui est l'indice pertinent pour discriminer /r/ de /l/ en anglais, sans qu'aucun entraînement explicite de catégorisation ou de feedback sur la qualité des réponses ne soit donné. Après un entraînement de seulement deux heures et demi sur cinq jours, les participants atteignent des performances généralement atteintes en deux à quatre semaines avec le paradigme classique de catégorisation à feedback (Iverson et al., 2005 ; Lively et al., 1994 ; Logan, Lively, & Pisoni, 1991). Cependant, il semblerait que cet entraînement réalisé sur des séquences synthétisées ne permette pas une généralisation optimale sur des paires minimales naturelles. En effet, malgré une augmentation des performances de 8.5%, la différence entre le pré-test et post-test n'est pas significative.

La méthode la plus efficace et de loin la plus écologique semble être la HVPT qui permet une amélioration significative des performances (même si celle-ci reste faible) ainsi qu'une importante généralisation à de nouveaux items. Cependant, on peut douter du fait que des entraînements en laboratoire améliorent significativement les performances dans le cadre d'une conversation, où le flux de parole du locuteur est loin de s'apparenter à celui des paires

[17] Trad. « Parole destinée à l'enfant»

minimales (Heeren & Schouten, 2008). Ces derniers ont en effet montré que le pic de sensitivité proche de la frontière catégorielle (c'est-à-dire le fait que les individus soient plus sensibles aux différences inter-catégorielles pour des exemplaires proches de la frontière catégorielle) n'est pas retrouvé chez les naïfs malgré le fait qu'ils parviennent à classifier un continuum de la même façon que les natifs après un court entraînement en laboratoire. Ce pic peut être obtenu mais cela demande bien plus d'expérience de la langue. Ils concluent que les améliorations des performances de classification observées en laboratoire « *should not be taken as evidence for (i) increased discrimination of the newly learned phonemes and (ii) learning of phoneme representations*[18] » (p. 2291).

1.3.3.3.2 IMMERSION

Un moyen somme toute plus écologique de tester les capacités d'adaptation des individus à de nouveaux contrastes phonologiques reste l'immersion en pays étranger. Dans ce sens, Logan, Lively et Pisoni (1991) testaient des japonais vivants aux Etats-Unis avec les mêmes stimuli, entraînement et mesures. La seule différence dans les résultats des deux études réside dans le fait que l'amélioration qu'ils avaient alors observée en termes d'amélioration pré-test/test était beaucoup moins importante (de l'ordre de 5-7%) alors même que les participants vivaient aux USA depuis 6 mois à 3 ans. Cependant, les performances au pré-test étaient ici plus importantes que dans l'étude de Lively et al. (1994) (65% pour Lively et al., 1994 qui testaient des Japonais au Japon contre 78 % pour Logan et al., 1991 qui testaient des japonais résidant aux Etats-Unis) attestant d'emblée de l'impact de l'immersion. Aoyama, Flege, Guion, Akahane-Yamada et Yamada (2004) ont également testé des Japonais vivant aux USA depuis environ six mois. Ces participants avaient tous appris auparavant l'anglais à l'école au Japon pendant environ six ans. Deux tests étaient effectués à un an d'intervalle, sans entraînement cette fois, afin de constater de l'amélioration de leur performance de discrimination des séquences CV. Même si la tendance est à l'augmentation, seule celle des enfants était significative et n'atteignait que 70% en score de discrimination contre 95% pour les contrôles. Les adultes n'améliorent donc pas leurs performances. Takagi & Mann (1995) ont testé des populations en immersion totale pendant au moins 12 ans dans un pays étranger avec une exposition intensive. Ils ont néanmoins constaté que cela ne fait que limiter les effets de la surdité phonologique. De la même façon, après une immersion d'environ sept mois en

[18] Trad. « ne devraient pas être considérées comme des preuves (i) d'amélioration de la discrimination des nouveaux phonèmes appris et (ii) d'un apprentissage de représentations phonémiques» .

Angleterre, des Danois ne parvenaient ni à percevoir, ni à produire la différence entre les /α/ et /Λ/ anglais (Hojen, 2003). En d'autres termes, il est difficile d'apprendre de nouvelles catégories phonologiques, même avec un entraînement, et les performances n'arrivent *quasi* jamais au même niveau que celles des natifs.

Cependant, certains individus n'ayant été exposés que tardivement à une L2 arrivent malgré tout à un niveau de maîtrise de la langue quasi similaire aux performances des natifs. C'est notamment le cas des participants de l'étude de Bongaerts et collaborateurs (1997). Des locuteurs natifs du néerlandais avec une excellente maîtrise de l'anglais ont été enregistrés lors de plusieurs tâches (e.g., parole spontanée, lecture, etc). Ces enregistrements, associés à ceux d'anglais natifs, ont été soumis à l'évaluation de juges anglophones. La plupart de ces apprenants tardifs avaient atteint un niveau jugé « *native-like* » par les juges. Mais les auteurs eux-mêmes constatent que de telles performances restent rares. En effet, nous ne sommes pas tous égaux face à l'apprentissage de nouveaux phonèmes.

Des questions restent donc ouvertes, portant notamment sur le fait que ces changements soient dus aux analyses auditives de bas niveau, à des changements d'ordre attentionnel, ou à des traitements de plus haut niveau.

1.3.3.3.3 CIRCUITS NEURONAUX IMPLIQUES DANS L'APPRENTISSAGE DE PHONEMES NON-NATIFS

Est-il possible de retrouver de la flexibilité neuronale pour la perception des sons non natifs? Une étude de Callan et al. (2004; 2003) nous éclaire sur les réseaux corticaux impliqués dans le processus d'identification des phonèmes non-natifs. Ils ont demandé à des japonais de distinguer le contraste de l'anglais /r/-/l/ qui est difficile pour eux et un contraste qu'ils possèdent dans leur langue (/b/-/g/). L'activité de leur cerveau était enregistrée via l'Imagerie par Résonnance Magnétique fonctionnelle (IRMf) avant et après la période d'un mois d'apprentissage des contrastes. Les résultats montrent des activations bilatérales différentes avant et après apprentissage dans des régions impliquées dans les traitements suivants (Figure 5): a) le traitement acoustico-phonétique du signal auditif (sulcus temporal supérieur (STS) antérieur gauche); b) les aspects articulatoires du signal auditif (gyrus temporal supérieur (STG) postérieur , planum temporale (PT, région sylvienne), et l'aire pariétale temporale supérieure (Stp)); c) la cartographie articulatoire-orosensorielle-auditive (Gyrus supra-marginal (SMG)); d) la planification de la production de la parole (aire de Broca, cortex pré-

moteur(PMC), insula antérieure); d) l'activation des modèles internes qui simulent les caractéristiques des entrées et des sorties du processus (cérebellum); e) la sélection de la valeur dépendante (ganglions de la base); et f) le contrôle de l'attention et la sélection [cortex cingulaire antérieur (ACC), cortex préfrontal dorso-latéral (DLPFC), aire motrice supplémentaire (SMA)). L'apprentissage d'une langue étrangère implique donc des modifications dans plusieurs réseaux corticaux.

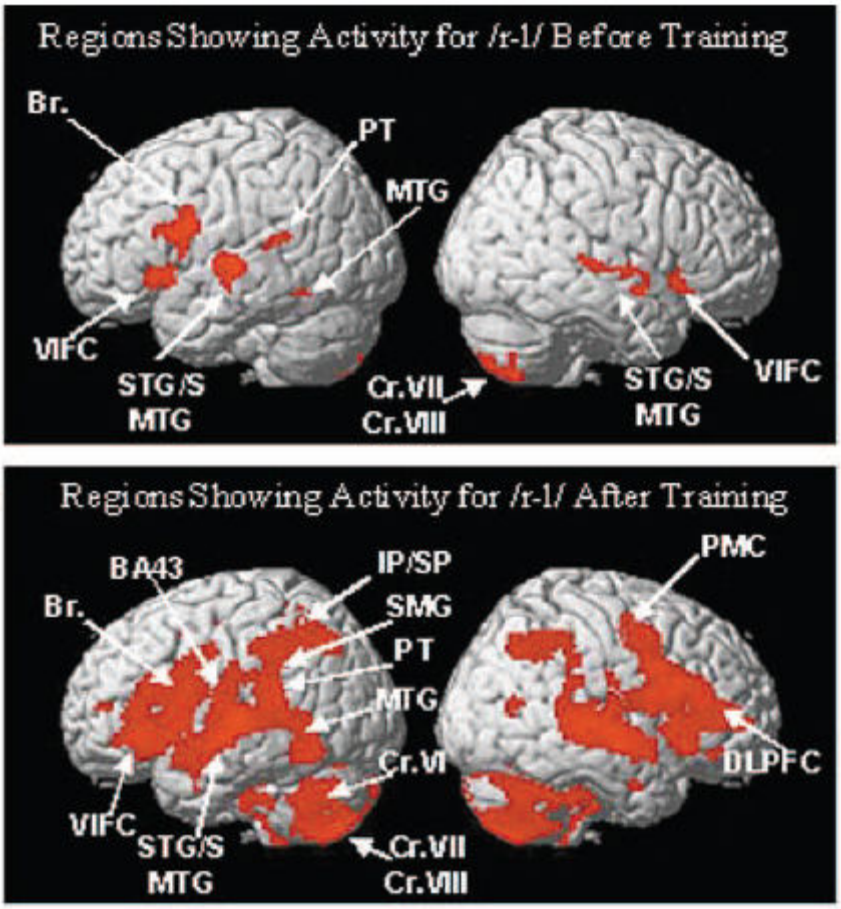

Figure 3. Modification des activations avant (encart supérieur) et après (encart inférieur) un apprentissage auditif sur le contraste /r/-/l/. (Extrait de Callan et al., 2003)

Les résultats montrent également des activations différentes et/ou additionnelles pour les contrastes difficiles. Selon les auteurs, pour identifier des phonèmes inexistants dans notre langue maternelle, nous ferions intervenir des réseaux corticaux impliqués dans la planification de la parole (Aire de Broca, insula antérieure, et PMC), ceux concernant les représentations auditives de la parole (acoustico–phonétique: STG; et articulatoire : STG postérieur, PT, STP), ainsi que ceux impliquant les représentations oro-sensorielles (SMG).

D'autre études, testant par exemple le contraste dental/rétroflexe de l'hindi ont également obtenu une activité plus intense du gyrus frontal inférieur et supra marginal en post-entraînement que avant celui-ci (Golestani & Zatorre, 2004). Plus récemment, Ventura-Campos et al. (2013) ont également observé des activations plus importantes dans le lobe pariétale supérieur gauche, le SMG gauche, l'opercule frontal inférieur et l'insula antérieur à la fois après l'apprentissage et durant le traitement de phonèmes non-natifs. L'implication de ces régions indique l'utilisation de cartes auditivo-articulatoires lors de l'identification de contrastes difficiles à percevoir. Ces zones sont généralement impliquées lorsque le contraste étudié diffère en terme de place d'articulation (Golestani & Zatorre, 2004).

L'acquisition d'un nouveau contraste passerait par deux processus différents en fonction du contexte de l'apprentissage. Celui-ci passerait par des feedbacks descendants dans le cas de l'apprentissage catégoriel, comme c'est le cas en laboratoire (Figure 6, encarre de gauche). Alors que les phonèmes natifs sont traités dans le gyrus temporal supérieur et la scissure temporale supérieure (ces deux structures sont en charge de l'acquisition des structures catégorielles natives), les traitements des sons qui en diffèrent sont redirigés vers les lobes frontaux. La sensibilité aux nouvelles catégories émergerait dans le Gyrus Frontal moyen, le Gyrus frontal inférieur gauche, le PMC et le cortex sensorimoteur gauche) (Myers & Swan, 2012). A force d'apprentissage, des retours aux zones temporales postérieures (STG et STS) seraient mis en place pour guider des changements de sensibilité à long terme. Ces feedbacks du frontal au temporal correspondent à la route d'apprentissage descendant. Le STG postérieur a d'ailleurs été identifié comme corrélant non seulement la sensibilité de la catégorisation des stimuli non-natifs appris mais également comme prédisant l'efficacité de l'apprentissage (Myers, 2014). Cependant, lors d'apprentissage lié à une exposition en milieu naturel (Figure 6, encarre de droite), l'adulte se baserait, comme l'enfant, sur les propriétés distributionnelles de la langue.

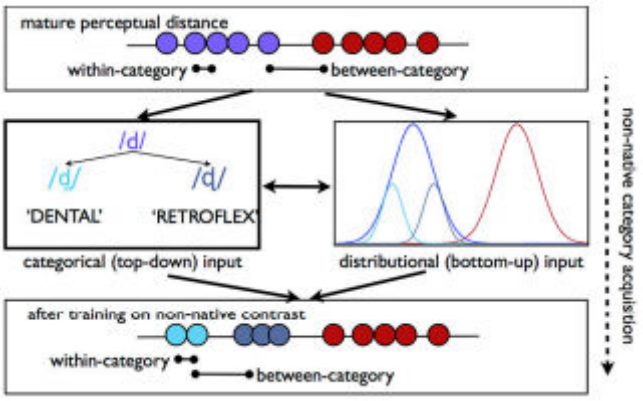

Figure 4. Schématisation des mécanismes mis en place lors de l'apprentissage (partie gauche) ou de l'acquisition (partie droite) de nouvelles catégories phonologiques. Exemple de la mise en place de catégories pour le phonème dentale et rétroflexe. (Extrait de Myers, 2014)

Dans ce cas, une voie ascendante serait empruntée afin de faire émerger de nouvelle catégorie. Cette seconde voie a été nettement moins étudiée que la première. En effet, la plupart des études s'intéressant à l'apprentissage de nouvelles catégories phonologiques sont réalisées avec un paradigme de catégorisation à feedback, privilégiant donc cette première voie. Les processus ascendants sont donc moins bien compris. Quelle que soit la route utilisée, l'apprentissage entraînera des modification des différences intra et inter catégories.

Certains réseaux semblent donc suractivés suite à l'apprentissage de nouveaux phonèmes, indiquant une flexibilité corticale qui permettrait d'apprendre des phonèmes non natifs. Cependant, et comme il l'a été dit précédemment, ces méthodes, comme les entraînements perceptifs, qui sont coûteux en temps, n'offrent pas nécessairement de grand pouvoir de généralisation des apprentissages à d'autres phonèmes ne partageant pas les mêmes traits discriminants.

1.3.3.4 FACTEURS DE VARIATION INTER-INDIVIDUELS

Nous avons tous constaté que les différences interindividuelles jouent un rôle importants lors de l'apprentissage des langues étrangères, notamment celles induites par des facteurs socio-culturels, cognitifs et maturationels (voir Dornyei, 2005) pour une revue sur l'impact des facteurs tels que les aptitudes langagières, l'autorégulation, les traits de personnalité et les styles cognitifs sur les aptitudes à acquérir une seconde langue). La motivation/implication émotionnelle lors de l'apprentissage, l'importance de l'exposition à la L2, le nombre de stratégies employées, voire également les capacités d'apprentissage qui varient d'un individu à l'autre impactent l'apprentissage d'une nouvelle langue (Flege, 1995 ; Flege, Bohn, & Jang,

1997 ; Moyer, 2007). Les capacités de mémoire phonologique ont elles aussi un rôle important pour l'acquisition de nouveaux contrastes. Service (1992) a mené une étude avec des enfants finlandais apprenant l'anglais. Elle suggère que la capacité à se représenter et de mémoriser des sons d'une langue joue un rôle important dans la capacité à apprendre une langue étrangère. Avant le début de l'apprentissage, les enfants devaient mémoriser et répéter des pseudo-mots finlandais, tâches impliquant toutes les deux la mémoire phonologique. Après deux ans d'apprentissage, ces mêmes enfants ont été testés sur leurs performances en anglais. Les résultats indiquent que les performances des enfants en anglais étaient corrélées avec les tâches de mémoire phonologique.

Il semblerait également que les individus diffèrent dans leur façon de catégoriser les phonèmes non-natifs. Polka, Colantonio et Sundara (2001) ont mené une étude dans laquelle des participants français devaient discriminer les phonèmes du contraste anglais /d/-/ð/. Pour ce même contraste, différents patterns d'assimilation ont été obtenus. Un participant les catégorisait en « Deux Catégories », quatre participants les classaient selon la « Catégorie Unique », et deux obtenaient un pattern que les auteurs ont classé en « Qualité/catégorie » (catégorisation selon le modèle PAM (Best, 1994, 1995) décrit au chapitre 1.3.3.5. Modèle d'assimilation perceptive). Cela vient du fait qu'à l'intérieur d'une même langue, même si les individus ont les mêmes catégories phonologiques, l'exemplaire prototypique de chaque catégorie n'est pas identique entre les individus induisant donc des différences de catégorisation lors de l'écoute de phonèmes non-natifs.

Des différences neuro-anatomiques ont également été observées entre les individus, ce qui pourrait expliquer éventuellement la facilité ou difficulté qu'on observe chez certaines personnes à apprendre des langues étrangères. Golestani, Molko, Dehaene, LeBihan et Pallier (2007) ont demandé à des francophones ne connaissant pas l'hindi de distinguer des consonnes dentales de consonnes rétroflexes de l'hindi. Ils ont constitué deux groupes d'individus sur la base de leurs scores dans une tâche de discrimination des consonnes insérées dans des syllabes. Un groupe comprenait les participants ayant appris rapidement à distinguer les consonnes et l'autre groupe, ceux qui ont appris plus lentement. Dans un deuxième temps, les auteurs ont mesuré les volumes de leur gyri de Heschl gauche et droit (localisés dans les gyri temporaux supérieurs). Le gyrus de Heschl abrite le cortex auditif primaire qui est la première aire corticale recevant les informations auditives issues de la périphérie. L'analyse des données anatomiques indique que les participants qui distinguaient le mieux les contrastes consonantiques de l'hindi avaient un cortex auditif gauche plus volumineux que ceux qui

avaient des performances moins bonnes dans cette tâche. Les différences entre les participants sont plus importantes pour le volume de matière blanche que de matière grise ce qui suggère, selon les auteurs, que les capacités à distinguer des contrastes phonologiques seraient liées soit à un plus grand nombre de fibres, soit à une meilleure myélinisation des fibres existantes ou les deux. Les données de cette recherche suggèrent donc que les individus qui ont des capacités importantes dans la distinction des sons d'une langue étrangère auraient des régions du cortex auditif primaire plus volumineuses. Cela permettrait une meilleure précision dans la représentation temporelle des sons, ce qui serait très utile pour discriminer des contrastes associés à des transitions acoustiques rapides, notamment celles qu'on observe dans beaucoup de consonnes.

Enfin, Sebastián-Gallés et al. (2012) ont également montré que les différences entre les individus ayant des difficultés pour acquérir un nouveau contraste et ceux ayant une facilité pour cette même acquisition se localisent, lors de l'écoute d'un contraste vocalique, non pas au niveau temporal (siège du cortex auditif primaire) mais au niveau du lobe frontal. Cela est cohérent avec le fait que des facteurs tels que le contrôle cognitif, notamment le contrôle de l'inhibition jouent un rôle dans la perception des phonèmes non-natifs chez les monolingues (Conboy, Sommerville, Jessica, & Kuhl, 2008), ces facultés étant gérées par le cortex pré-frontal inférieur (Aron, Robbins, & Poldrack, 2004).

1.4 CONCLUSION

La surdité phonologique, bien que robuste, ne reflète pas une perte de sensibilité totale, mais seulement une restriction du champ sensoriel, notamment due à l'impact des catégories natives. Il semblerait que ce soit au travers de ces catégories natives que nous percevons les phonèmes non-natifs, entrainant des assimilations phonologiques qui biaisent notre perception. Avant 12 ans, ces catégories sont encore assez flexibles pour permettre une acquisition efficace des phonèmes qui n'appartiennent pas au répertoire. Mais plus tard, ces capacités se figent avec les catégories phonologiques. Les effets de la surdité phonologique semblent cependant modifiables à plus ou moins long terme par une immersion ou des entraînements, pour des résultats qui n'atteignent cependant que rarement les performances des natifs.

Existerait-il un moyen plus automatique et à court terme qui permettrait d'utiliser des indices pertinents afin de désambiguïser le signal acoustique ? Des études récentes montreraient que la présentation des informations de nature articulatoire pourrait améliorer la perception de certains phonèmes (Davis & Kim, 1998 ; Navarra & Soto-Faraco, 2007). En effet cette information, essentiellement visuelle, est utilisée de manière automatique lors de conversation face-à-face et pourrait fournir des indices importants dans le cadre de la perception d'une langue non-native. Explorer cette hypothèse constitue un des objectifs de cette thèse.

CHAPITRE 2

PERCEPTION DE LA PAROLE AUDIOVISUELLE

«Speech perceivers are informational omnivores» (2004, p.189)

— Carol Fowler

Lorsque nous tenons une conversation face-à-face, nous sommes généralement exposés à plusieurs flux auditifs. En plus du discours émis par notre interlocuteur, nous percevons généralement de nombreux bruits ambiants, que ce soit les paroles d'autres personnes autour de nous ou simplement le bruit de la circulation ou de la télévision. Malgré ce brouhaha ambiant, notre compréhension n'est la plupart du temps pas altérée. Cela est notamment dû à notre faculté, lors d'une conversation face-à-face, à utiliser continuellement et inconsciemment les gestes oro-faciaux du locuteur, même si les informations auditives sont souvent suffisantes pour comprendre la parole.

Comment utilisons-nous ces informations et quels avantages nous donnent-t-elles pour percevoir la parole ? Ce sont les questions auxquelles nous tenterons de répondre dans ce chapitre en nous intéressant en premier lieu à la perception de la langue maternelle, pour nous diriger par la suite vers la perception audiovisuelle des phonèmes des langues étrangères.

2.1 PERCEPTION AUDIOVISUELLE DE LA LANGUE MATERNELLE

2.1.1 GENERALITES

Lorsque nous tenons une conversation dans des conditions naturelles, nous avons toujours un contact visuel avec notre interlocuteur et nous pouvons ainsi tirer parti de nombreuses sources d'informations pour optimiser la compréhension de la parole (e.g., posture, mouvements qui accompagnent le discours, expression du visage, gestes articulatoires visibles, etc). Cette capacité à lier deux sources d'information se révèle d'ailleurs cruciale pour l'émergence d'une représentation unifiée du monde qui nous entoure (Stein & Meredith, 1993). Cela inclue l'aptitude à percevoir les attributs visuels et auditifs de la parole comme un événement unique.

2.1.1.1 CHEZ LE BEBE

Cette faculté de lier un évènement visuel à sa contrepartie auditive émergerait à la naissance ou assez précocement dans le développement (Gibson, 1969 ; Thelen & Smith, 1994). Même en dehors du cadre langagier, les nouveaux nés de deux jours ont déjà des habiletés perceptives inter-sensorielles qui leur permettent de lier des événements audiovisuels temporellement synchrones (Slater, Quinn, Brown, & Hayes, 1999). Expérimentalement, dans le cadre de la perception du langage, cette capacité de faire correspondre un son à sa réalisation articulatoire s'étudie via la présentation simultanée de deux visages côte à côte articulant silencieusement deux phonèmes/syllabes/phrases. Une présentation sonore (congruente ou incongruente) est associée aux vidéos. Celle-ci correspond à la réalisation articulatoire de l'une ou l'autre des vidéos. Si les enfants sont capables de faire correspondre un son à son articulation, ils regarderont plus longtemps la vidéo correspondant au son entendu. Cette méthodologie a permis de mettre en évidence que les bébés savent dès 4-5 mois détecter l'asynchronie de la parole (Lewkowicz, 2010) mais également associer un « eeeee » auditif à la réalisation articulatoire qui convient (Kuhl & Meltzoff, 1982 ; Kuhl, Williams, & Meltzoff, 1991 ; Patterson & Werker, 2003 ; Patterson, 1999 pour d'autres résultats sur des voyelles isolées). La même aptitude a également été observée lors de la perception de consonnes. A six mois, les enfants hispanophones et anglophones testés par Pons, Lewkowicz, Soto-Faraco et Sebastián-Gallés (2009) parviennent à associer un /ba/ ou

un /va/ auditif à la réalisation articulatoire qui convient. Weikum et al. (2007) parviennent aux mêmes résultats en utilisant des phrases.

Cette prise en compte de la partie visible de la parole est cruciale. Elle nous donne accès à une grande partie des caractéristiques articulatoires (Gentil, 1981 ; Jiang, 2003 ; Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998 ; Summerfield, 1987) qui permettent de faciliter l'apprentissage de certaines catégories phonologiques (Teinonen, Aslin, Alku, & Csibra, 2008). Une étude de Mills (1987) atteste que les informations visuelles sont critiques pour l'acquisition de certain contrastes, puisque les enfants aveugles ont par exemple des difficultés à acquérir les contrastes /m/-/n/ ou /b/-/k/ qui sont décrits comme « easy-to-see » et « hard-to-hear[19]» (voir aussi Dodd, McIntosh, & Woodhouse, 1998). La prise en compte des informations provenant de différentes modalités sensorielles est donc essentielle lors du développement.

2.1.1.2 CHEZ L'ENFANT

Alors que ces capacités d'association multimodale s'affinent au cours du développement, permettant des associations de plus en plus complexes, la courbe développementale de la perception multimodale n'est cependant pas incrémentale. Même si les capacités à lier perceptivement deux évènements vont en s'améliorant, ce n'est pas pour autant qu'elles resteront efficaces plus tard dans l'enfance. En effet, des études utilisant l'effet McGurk (McGurk & MacDonald, 1976) attestent que les capacités des enfants à utiliser les informations visuelles changent au cours du développement. L'effet McGurk est obtenu avec la présentation d'un /ga/ visuel associé à un /ba/ auditif. Il se base donc sur la présentation de stimuli incongruents. Le percept qui en résulte n'est ni /ga/ ni /ba/ mais /da/ ou /ða/. Ces résultats reflètent une illusion perceptive, lié à la fusion des deux informations. Alors que Burnham & Dodd en 2004 montrent qu'à 5 mois, les bébés sont sensibles à l'effet McGurk, les résultats de McGurk & MacDonald (1976) et de Sekiyama & Burnham (2008) vont dans le sens d'une diminution de l'effet McGurk chez des enfants plus âgés. Cela reflète selon eux une diminution de l'intégration bimodale chez les enfants. Selon Massaro, Thompson, Barron et Laren (1986), la diminution de l'effet McGurk serait due à une sous exploitation de l'information visuelle. Dans cette étude, des enfants et des adultes devaient identifier les syllabes /da/ et /ba/, présentées soit en modalité visuelle seule, soit en présentation bimodale.

[19] Trad. "facile à voir" "difficile à entendre"

Les stimuli auditifs utilisés lors de la présentation audiovisuelle avaient été synthétisés (i.e., manipulation de F1, F2 et F3) pour obtenir un continuum de 5 stimuli allant d'un /ba/ à un /da/, afin de générer de l'ambiguïté. Chaque exemplaire auditif a été présenté avec l'articulation visible du /ba/ ou du /da/. Leurs résultats (Figure 7) montrent une influence bien plus marquée de l'information visuelle chez les adultes.

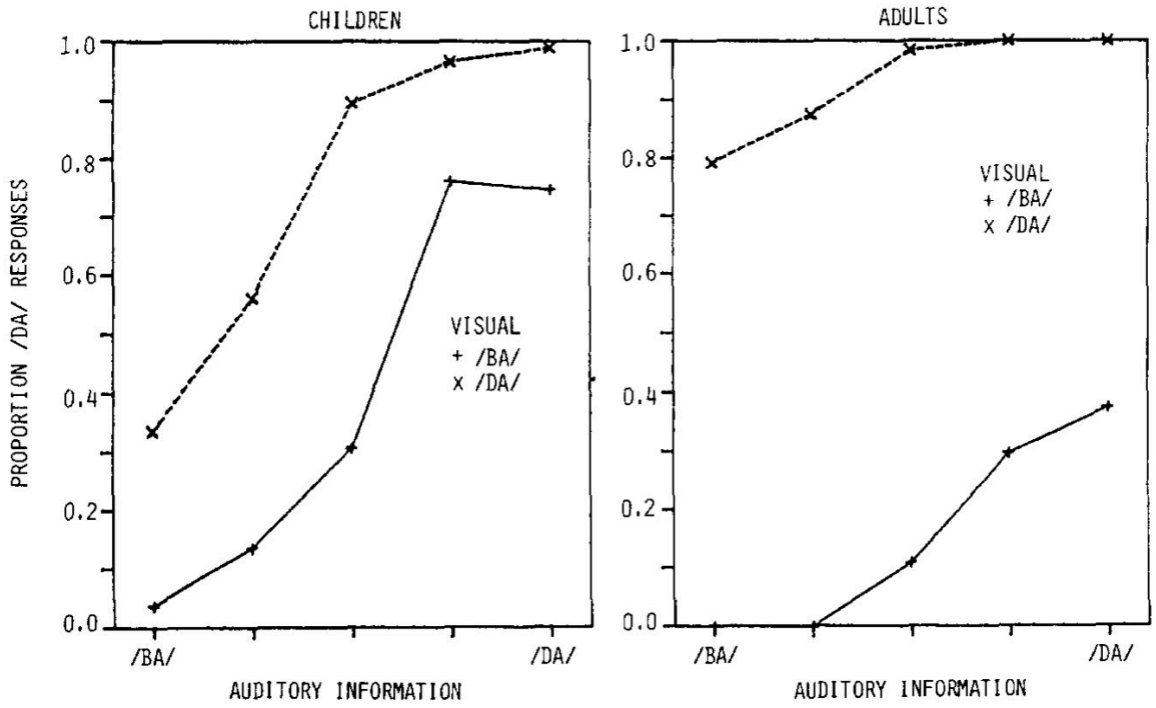

Figure 5. Proportion de réponses /da/ en fonction du niveau d'information auditive et visuelle dans la condition de présentation bimodale. (, Extrait de Massaro, Thompson, Barron et Laren, 1986).

En effet, les adultes donnent une majorité de réponses /da/ quelque soit le signal auditif lorsque l'articulation visible est celle d'un /d/ et *vice versa* pour le /ba/. Pour les enfants cependant, on remarque que le nombre de réponses /da/ est important lorsque le signal auditif est plus proche d'un /d/ indépendamment de l'articulation visible. Cela n'est pour autant pas dû à une surexploitation des informations auditives par les enfants car aucune différence n'était observée entre les groupes en condition audio seule dans une précédente étude (Massaro, 1984). Cela s'expliquerait donc bien par des aptitudes moins importantes en lecture labiale comme l'attestent les résultats lors de la présentation visuelle seule (i.e., 96 % de réponses correctes pour les adultes contre 79% chez les enfants). Il semblerait que la capacité à lire sur les lèvres demande un certain temps pour se stabiliser. Cette difficulté

s'étendrait en général jusqu'a 6-10 ans (Massaro et al., 1986 ; Sekiyama & Burnham, 2004). A l'âge adulte, lorsque les capacités à utiliser l'information labiale sont stabilisées, nous tirons de nombreux avantages de la présentation audiovisuelle.

2.1.1.3 CHEZ L'ADULTE

2.1.1.3.1 PRESENTATION AUDIOVISUELLE DANS LE BRUIT

La présentation bimodale permet d'améliorer la perception de la parole dans de nombreuses conditions. Avoir accès aux informations visuelles permet d'augmenter l'intelligibilité des phonèmes ainsi que la compréhension du discours quand les conditions d'écoute ne sont pas adéquates (Erber, 1969 ; MacLeod & Summerfield, 1987 ; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). Ce sont Sumby & Pollack (1954) qui ont en premier montré l'impact positif de la présentation audiovisuelle lorsque les conditions d'écoute sont bruités ou *adverses*. Dans cette étude, des mots bisyllabiques étaient présentés en modalité auditive et audiovisuelle. Le signal acoustique était dégradé à différents niveaux de signal/bruit[20]. L'analyse des scores d'identification de mots étaient meilleurs lorsque la présentation était bimodale qu'en présentation auditive seule. De plus, ils ont mis en évidence que l'avantage lié à la présentation audiovisuelle augmente d'autant plus que le signal est bruité. Ils ont donc fourni une preuve de l'utilisation des mouvements oro-faciaux par des auditeurs normaux-entendants dans des conditions adverses. Ces résultats ont par la suite été répliqués par Binnie, Montgomery et Jackson (1974), Erber (1969) et Middelweerd & Plomp (1987) en anglais, puis par Benoît, Mohamadi et Kandel (1994) pour le français. Binnie et al. (1974) ont notamment mis en évidence que l'avantage de la présentation bimodale par rapport à la présentation des informations auditives bruitées seules était lié au fait que les informations visuelles permettent de réduire significativement les confusions entre des séquences CV dont la place d'articulation diffère.

Ces premières études ont également permis de quantifier l'amélioration relative de la perception du signal auditif lors d'une présentation audiovisuelle, par la soustraction des scores obtenus lors de la présentation auditive aux scores obtenus en modalité audiovisuelle.

[20] Le rapport signal sur bruit désigne le rapport entre la puissance du signal auditif et la puissance du signal parasite. Cela permet de quantifier le niveau de bruit inséré dans le signal acoustique d'intérêt. Plus le rapport signal sur bruit est grand, plus le bruit est faible par rapport au signal et plus le signal est perceptible. Ainsi un rapport signal sur bruit de -18dB masquera bien plus le signal acoustique qu'un signal rapport sur bruit de -6dB.

Ainsi, Sumby & Pollack (1954) ont mis en évidence qu'une présentation audiovisuelle correspond à une amélioration du rapport signal/bruit de 15dB par rapport à une présentation audio seule. Calbour et Dumont (2002) ainsi que MacLeod et Summerfield (1987) l'ont également estimé et notent un bénéfice moyen de 11 dB dû à l'ajout des informations visuelles. Cela entraîne un confort de perception supérieur à l'écoute seule. Plus tard, il a été mis en évidence qu'en fait ce sont les niveaux de bruit intermédiaires qui permettent la meilleure amélioration (Ross et al., 2007).

La présentation audiovisuelle permet d'augmenter le taux d'identification correct de voyelles même lorsqu'elles ne sont pas visuellement distinguables. Les participants de Schwartz, Berthommier et Savariaux (2004) devaient identifier /y/ et /u/ présentés seuls ou dans une syllabe avec un rapport signal/bruit de -9 dB. Ils parvenaient en effet à identifier plus efficacement /ty/ que /tu/ lors d'une présentation audiovisuelle par rapport à la présentation auditive alors que l'articulation visible de ces deux voyelles est la même. Enfin, les informations visuelles permettent, toujours lorsque les conditions d'écoute sont bruitées, des activations lexicales plus rapides dans des paradigmes d'amorçage (Fort et al., 2012) et contribuent au processus de reconnaissance lexicale (Fort et al., 2012).

2.1.1.3.2 PRESENTATION AUDIOVISUELLE SANS BRUIT

2.1.1.3.2.1 STIMULI NON-CONGRUENTS

L'information visuelle permet donc de mieux comprendre notre interlocuteur malgré le bruit ambiant. Il semblerait cependant que la prise en compte de ces informations soit automatique, le bruit n'étant pas une condition nécessaire pour leur utilisation. En effet, le fameux « effet McGurk » (Jesse & Massaro, 2010) n'implique pas l'ajout de bruit au signal, mais se base sur la présentation audiovisuelle de deux informations qui sont en conflit. Il reste un des exemples les plus connus et les plus répliqués de l'intégration automatique des informations auditives et visuelles, même sans ajout de bruit pour « forcer » l'utilisation des informations visuelles. D'ailleurs, nul besoin de mettre en jeu un stimulus aussi complexe que la parole car cette capacité à unifier un évènement audiovisuel s'observe également pour des sons, comme ceux des instruments de musique (Saldaña & Rosenblum, 1993)[21].

[21]Expérience intégrant 2 types de jeux au violoncelle (continuum sonore de « cordes pincées » à « cordes frottées ») au paradigme de McGurck & MacDonald. Associées aux images congruentes ou non d'un

2.1.1.3.2.2 STIMULI CONGRUENTS

Même dans le cas de stimuli congruents, et quand les conditions d'écoute sont normales (i.e., non bruitées), la présentation audiovisuelle apporte des avantages, en donnant par exemple une impression d'augmentation du volume sonore ce qui améliore l'intelligibilité (Saldaña & Rosenblum, 1993) cité par Schwartz et al., 2004). Ces auteurs ont également noté que le seuil auditif diminue de 1-2 dB lorsque le son est accompagné des mouvements labiaux. Ces derniers permettent également de mieux comprendre un discours exprimé avec un fort accent étranger (Burnham, 1998). Des études ont mis en évidence que ce bénéfice était observé aussi bien en situation conversationnelle (Cerrato, Leoni, & Falcone, 1998), que lors de la perception de logatomes VCVCV (Benoît et al., 1994).

L'information visuelle est donc largement utilisée pour améliorer la compréhension de la parole ainsi que des unités sous lexicales qui la composent, aussi bien en présence de bruit que lorsque les conditions d'écoute sont bonnes. Elle fournirait des informations complémentaires au signal auditif, ce qui permettrait de le désambiguïser, voire de faciliter les activations lexicales. Enfin, son utilisation se révèle être automatique.

2.1.1.4 LES VISEMES

Mais alors, que percevons-nous lorsque nous regardons notre interlocuteur s'exprimer ? Des *visèmes* (Fisher, 1968) ! Pendants visuels du phonème[22]. Un visème est la plus petite unité contrastive de l'articulation visuelle. Calbour et Dumont (2002) les décrivent comme « les phonèmes dont l'articulation visible procède du même geste facial » (p. 11). Lorsque nous parlons, nous modifions les caractéristiques de notre conduit vocal en mettant en action séquentiellement et/ou simultanément différents articulateurs et résonateurs. Ainsi, nous modelons le signal acoustique de sortie. Cela engage de nombreux mouvements articulatoires qui sont plus ou moins visibles.

Il est d'ailleurs important de noter que les lèvres ne sont pas les seuls vecteurs d'information lorsque nous observons quelqu'un parler. Même si les deux tiers de l'information utilisée lors de présentation d'un visage entier est fournie par les lèvres (Benoît, Guiard-Marigny, Le Goff, & Adjoudani, 1996), les dents (McGrath, 1985 cité par

violoncelliste pinçant ou frottant les cordes, la même intégration que celle observée avec des syllabes a été obtenue.

[22]D'où son nom tiré de la contraction des mots anglais « ***vis***ual phon***eme*** » (Léon, 1992).

Summerfield Quentin, 1991 ; Thomas & Jordan, 2004), la mâchoire (Guiard-Marigny, Tsingos, Adjoudani, Benoît, & Gascuel, 1996 ; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998) et la langue (Badin, Tarabalka, Elisei, & Bailly, 2010), constituent également des indices importants. Les zones qui ne sont pas directement associées à la région buccale amènent également un avantage. C'est le cas des joues (Badin, Tarabalka, Elisei, & Bailly, 2010) ainsi que le haut de la tête (Preminger, Lin, Payen, & Levitt, 1998) qui transmettent également des indices relatifs à la parole. Cela explique pourquoi les performances en lecture labiale s'améliorent lorsque l'interlocuteur a accès aux informations fournies par le visage entier par rapport à la zone buccale seule (Thomas & Jordan, 2004). Il a également été montré que les mouvements de la tête fournissent des informations linguistiques exploitables, notamment sur l'intonation (Thomas & Jordan, 2004). Ce bénéfice supplémentaire suggère que même si les mouvements labiaux constituent un moyen efficace de véhiculer des informations visuelles sur la production de la parole, d'autres indices oro-faciaux situés autour de cette zone sont également exploitables.

L'être humain se révèle être naturellement un assez bon labio-lecteur. Il a été évalué que l'information fournie par le discours visible permet de discriminer 40 à 60 % des phonèmes d'une langue et de 10 à 20 % des mots (Bernstein, Iverson, & Auer, 1997 cité par Schwartz, 2011). Cependant, cela varie bien évidemment en fonction de la langue étudiée. En anglais, le taux de réussite lors de l'identification labiale (de syllabes, mots isolés ou mots dans une phrase porteuse) est de 40 à 60 % en moyenne pour les normo-entendants et de 60-80 % pour les sourds (Bernstein et al., 1997). Auer & Bernstein (1997) ont montré que 54 % des mots monosyllabiques fréquents de l'anglais pouvaient être identifiés sur la seule base des informations visuelles grâce à 12 visèmes (voir aussi Auer, 2002). Iverson, Bernstein et Auer, (1998) ont également observé que les mots de plus d'une syllabe étaient plus facilement identifiables. Ils comportent effectivement plus d'information que des mots monosyllabiques qui peuvent être sujets à plus de confusion (15 % des monosyllabiques sont reconnus contre 75 % des multi-syllabiques). Globalement, nous sommes tous capables de distinguer « mère » (/mɛʁ/) de « nerf » (/nɛʁ/) avec les seuls indices visuels. Cette distinction se fait grâce au place d'articulation (Figure 8), car l'occlusive bilabiale /m/ est classée dans une catégorie visémique distincte de l'occlusive alvéolaire /n/.

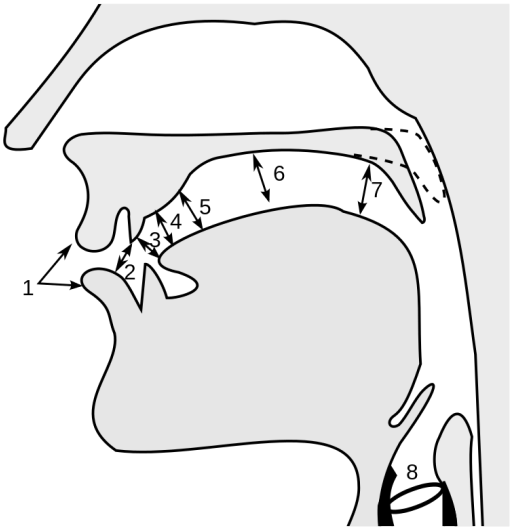

Figure 6. Schéma des différentes places d'articulation utilisées pour la réalisation des consonnes. 1 : bilabiale, 2 : labiodentale, 3 : dentale, 4 : alvéolaire, 5 : post-alvéolaire, 6 : palatale, 7 : vélaire, 8 : glottale.

Ces classes de phonèmes visuellement distincts étaient déjà envisagées dans les années 30 par Deland (cité par Fisher, 1968). Nous allons dès à présent détailler ces classes en nous intéressant dans un premier temps aux consonnes pour nous pencher par la suite sur les voyelles.

2.1.1.4.1 CARACTERISATION DES VISEMES CONSONANTIQUES

Les classes visémiques des consonnes ont été largement étudiées. Le nombre de visèmes répertoriés s'échelonne de trois à onze en fonction des études (Binnie et al., 1974 ; Heider & Heider, 1940 ; Massaro, Cohen, Gesi, & Heredia, 1993). Les variations observées sont en majorité dues à la population étudiée (i.e normaux-entendants ou sourds), aux conditions de présentation (i.e., visuelle seule ou audiovisuelle, avec ou sans bruit, etc), à la langue testée, au locuteur, mais également à la position de la consonne dans la séquence. Les catégories visémiques identifiables de la langue française varient en fonction de la position de la consonne dans la séquence : initiale, médiane ou finale. Gentil, (1981 cité par Cathiard, 1994) les identifie comme suit :

Sur les seize consonnes du français, six catégories visuelles se démarquent :

- Les occlusives bilabiales /p b m/
- Les fricatives labiodentales /f v/
- Les fricatives palatales protruses /ʃ ʒ/

Ces classes sont toujours visibles indépendamment de leur place dans la séquence. Les classes suivantes sont quant à elles visibles uniquement lorsqu'elles sont situées à la fin de la séquence :

- Les occlusives dentales /t d n/
- Les constrictives alvéolaires /s z/
- Les liquides /l ʁ/

Les phonèmes /k g ɲ/ ne sont ici pas considérés car leur place d'articulation, très en arrière du conduit vocal, ne permet pas (dans la théorie) de les considérer comme des classes visémiques à part entière.

On constate que plusieurs phonèmes sont regroupés dans la même catégorie visémique, ce sont des sosies labiaux, également appelés homophènes (Cathiard, 1994), c'est-à-dire que même s'ils sont différents acoustiquement, ils sont souvent confondus en modalité visuelle seule, car les mouvements articulatoires liés à leur prononciation sont similaires, comme lorsque l'on prononce /p/ et /b/.

Les phonèmes consonantiques les plus facilement distingués correspondent à ceux qui sont articulés à l'avant du conduit vocal (i.e., /p b m/ et /f v/) et ceux qui disposent d'une caractéristique articulatoire hautement visible comme une protrusion des lèvres (i.e., /ʃ ʒ/). L'aperture labiale, qui est une caractéristique vocalique, permet elle aussi de distinguer les consonnes bilabiales des labiodentales. Cependant, et comme le soulignent Lubker & Gay, (1982), le nombre de consonnes bilabiales demeure relativement restreint (que ce soit en français ou dans la plupart des langues du monde). En ce qui concerne la classe renfermant /p b m/, Abel, Barbosa, Black, Mayer et Vatikiotis-Bateson (2011) soulignent que la définition de cette catégorie, comme d'autres, se base sur le pourcentage de confusion intra-catégorie de ces phonèmes, c'est-à-dire le nombre de confusion à l'intérieur d'une classe visémique donnée. Lorsque ce pourcentage s'élève à 70-75% (Walden, Prosek, Montgomery, Scherr, & Jones, 1977) les auteurs considèrent que ces phonèmes font partie de la même catégorie visémique. Cependant, et même s'ils sont difficilement discriminables les uns des autres, il est possible de les distinguer. Dans leur étude, Walden et al. (1977) montrent en effet que /m/ et /p/ sont correctement identifiés au-dessus du hasard lorsque le locuteur est vu de face ou de profil. Il est important de signaler que l'arrondissement labial qui est spécifique de la réalisation articulatoire de certaines voyelles est également observé lors de la production de certaines consonnes. En français, les consonnes fricatives post-alvéolaires /ʃ ʒ/ sont de fait réalisées avec un arrondissement labial (Abry & Boë, 1980; Descout, Boë, & Abry, 1980).

Les autres consonnes pourront être labialisées mais seulement par influence d'un contexte vocalique arrondi. Inversement, les phonèmes consonantiques les plus facilement confondus sont articulés au niveau de ou derrière les dents, c'est-à-dire plus en arrière du conduit vocal (i.e., /s z t d n k g ŋ ʁ/). Ce constat renvoie à la notion de saillance perceptive (Summerfield, 1987).

2.1.1.4.2 CARACTERISATION DES VISEMES VOCALIQUES

En modalité auditive, les voyelles se distinguent principalement grâce aux valeurs du premier et du deuxième formant. Au niveau visuel, parmi les trois dimensions articulatoires qui permettent de différencier les voyelles de toutes les langues (i.e., aperture, labialité et position de la langue) seules deux sont visibles aux lèvres : la différence d'ouverture labiale (verticale) qui permet par exemple de distinguer /a/ et /i/ et la dimension d'arrondissement-étirement qui permet de différencier la voyelle arrondie comme /y/ de la voyelle non-arrondie ou étirée /i/ (Figure 9). Ce sont bien entendu les voyelles de l'extrémité du triangle vocalique qui sont les plus facilement identifiables : le /a/ a la plus grande aperture, le /i/ est la plus étirée, le /u/ est la plus arrondie (Gentil, 1981).

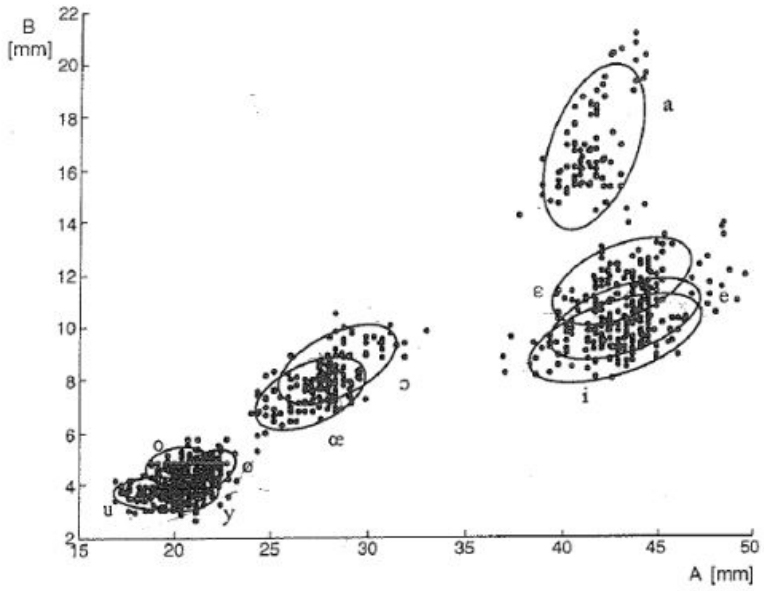

Figure 7. Représentation des voyelles du français dans le plan des paramètres géométriques de hauteur (B) et d'arrondissement-étirement (A). (Tiré de Robert-Ribes et al., 1998)

Dans les langues comme le français, la première étude portant sur les visèmes a été réalisée par Mourand-Dornier (1980). Le trait le plus informatif visuellement est la *labialité*. Elle

permet de scinder en deux catégories le répertoire vocalique avec d'une part, les voyelles arrondies /y u o õ ø ɔ œ/, et d'autre part les voyelles non-arrondies comme /i a e ɛ ɛ̃/. C'est à l'intérieur de ces deux catégories que les confusions sont les plus fréquentes (Cathiard, 1994). A l'intérieur de chacune, les voyelles diffèrent sur leur *aperture buccale* et sur *l'étirement*. Plus tard, Robert-Ribes et al. (1998) ont effectué un test perceptif en modalité visuelle seule sur les voyelles /a e i o u y ø/ pour voir si ces indices visuels étaient exploités pour l'identification des phonèmes. Ils ont mis en évidence trois visèmes : /a/, /e i/ et /o u y ø/. L'arrondissement des lèvres (paramètre , Figure 9) semble être la dimension visuelle la plus pertinente et permet de distinguer les voyelles arrondies des non-arrondies (e.g., distinguant un /y/ d'un /i/). La hauteur (paramètre B, Figure 9) permet quant à elle de subdiviser les deux premières catégories (e.g., distinguant un /a/ d'un /i/ et d'un /y/). L'arrondissement des lèvres et la hauteur semblent donc être des traits pertinents pour la discrimination perceptive des voyelles en modalité visuelle seule.

Pour les seize phonèmes vocaliques, seules sept catégories sont visuellement distinctes :

- Le visème /a ã/
- Le visème /o õ ø/
- Le visème /y u/
- Le visème /ɛ ɛ̃/
- Le visème /e/
- Le visème /i/
- Le visème /œ ɔ/

Comme pour les visèmes consonantiques, il existe des sosies labiaux. Cependant, il est à noter que contrairement aux consonnes qui sont séparées par d'importantes variations, les voyelles sont non seulement plus proches au niveau de leur réalisation articulatoire mais diffèrent également beaucoup d'une langue à l'autre, ce qui explique que certains auteurs relèvent moins de visèmes vocaliques que d'autres (Istria, Nicolas-Jeantoux, & Tamboise, 1982). Cathiard (1994) ajoute même que « certaines frontières de décision [entre des visèmes vocaliques] sont des frontières très fines puisque se jouant parfois sur moins d'un millimètre » (p.151).

2.1.1.5 FACTEURS IMPACTANT L'UTILISATION DES VISEMES

De nombreux facteurs sont connus pour influencer le repérage des différentes informations et constituent ainsi des limites à la lecture labiale. Les phénomènes de saillance, coarticulation et de stabilité sont des éléments majeurs entrant en jeu dans la lecture labiale.

2.1.1.5.1 IMPACT DE LA SAILLANCE ET DE LA COARTICULATION

La *saillance* (Summerfield, 1987) est un des facteurs principaux qui module l'utilisation des informations labiales. Un élément est dit saillant lorsque celui-ci s'impose à l'attention, sans effort particulier et qu'il apparaît facilement au regard du contexte. Par exemple, Giraud & Poeppel (2012) expliquent que les syllabes qui commencent par une consonne bilabiale entraînent plus de facilitation que celles qui commencent par une vélaire car elle est constituée d'indices visuels plus évidents/visibles. Certains visèmes consonantiques (et également vocaliques) sont plus facilement reconnaissables que d'autres parce que plus visibles. Ces sont les consonnes articulées à l'avant du conduit vocal qui présentent le plus d'avantages lors d'une présentation audiovisuelle car elles sont facilement repérables pour le labiolecteur (i.e., les bilabiales, les protruses ou les labiodentales ; Cathiard, 1994). Cependant, replacée dans une démarche expérimentale, et comme le notent Paris, Kim et Davis (2013a), il faut reconsidérer la façon dont est définie la saillance visuelle. Celle-ci est non seulement déterminée par les caractéristiques articulatoires mais également par la tâche qu'effectue le participant. En effet, la saillance est relative aux autres visèmes considérés. Les tâches, notamment celle d'identification à choix forcé, et qui plus est lorsque le choix des réponses est restreint, entraînent une diminution de la saillance. Prenons comme exemple une tâche d'identification de *deux* consonnes inscrites dans des syllabes (e.g., /ba/ et /da/). Jesse & Massaro (2010) ont montré qu'en termes de caractéristiques articulatoires, une consonne bilabiale (e.g., /ba/) contient trois fois plus d'information qu'une consonne sans fermeture complète (e.g., /da/), elle est donc plus saillante. Cependant, dans le cadre d'une tâche à deux choix forcés, ces caractéristiques n'ont plus le même poids. En effet, si le participant ne perçoit pas les indices indiquant un /ba/ (fermeture complète des lèvres), il ne lui reste d'autre choix que de répondre /da/ (ou *vice versa*). En d'autre termes, et comme le disent Paris, Kim et Davis, « *although the [ba] token was more salient than the [da] one, there were only two response options in this experiment (ba or da), meaning any difference in their salience was*

minimized (i.e., if the cue for a [ba] was not present, a [da] could be presumed)[23] » (2013, p. 356). Ainsi, dans le cas de réponse à choix forcé, les deux stimuli sont aussi saillants et faciles à détecter l'un que l'autre, indépendamment de caractéristiques articulatoires propres.

Tout comme pour la perception auditive des phonèmes, les indices visuels sont également modulés par la *coarticulation*. En effet, la présence de phonèmes adjacents modifie les réalisations articulatoires canoniques et influence donc leur identification/discrimination en modalité visuelle (Benguerel & Pichora-Fuller, 1982 ; Benoît et al., 1994 ; Gentil, 1981). Les occlusives bilabiales, les fricatives labiodentales et les fricatives palatales protruses peuvent ainsi être placées dans la catégorie des consonnes stables (Istria et al., 1982). Elles sont donc « facilement » identifiables, et ce de manière assez constante, quelle que soit leur place dans le mot. Les occlusives dentales, les liquides et les fricatives alvéolaires (i.e., /t d n l s z/) peuvent être considérées comme « variables », c'est-à-dire que leur facilité d'identification est fonction de leur place dans le mot et de leur voisinage phonémique. Par exemple, l'articulation du /t/ est très difficilement perceptible dans « tiroir » du fait de l'étirement du /i/ alors qu'elle l'est plus dans « tapis » du fait de l'aperture du /a/. Enfin, les occlusives vélaires et la vibrante vélaire (i.e., /k g ʁ/) sont considérées comme invisibles, elles sont donc très mal perçues en modalité visuelle seule, peu importe les effets de coarticulation.

De nombreuses études ont été consacrées à l'observation de la coordination entre voyelles et consonnes dans la production de séquences VCV et notamment sur les effets de la hauteur imposée par la voyelle sur les consonnes adjacentes (Benguerel & Pichora-Fuller, 1982 ; Benoît et al., 1994 ; Gentil, 1981). En 1994, Benoît, Mohamadi et Kandel ont étudié l'identification des consonnes /b v ʁ l z ʒ/ et des voyelles /i a y/ insérées dans des séquences de type /VCVCV/. Ils ont pu montrer qu'en présentation visuelle, le contexte /a/ était le plus favorable à l'identification de consonnes, suivi du /i/ puis du /y/. Les caractéristiques articulatoires des phonèmes adjacents modifient donc leur intelligibilité (voir Gentil, 1981 et Massaro et al., 1993 pour des résultats similaires en français et en anglais, respectivement). Cela est dû aux contraintes articulatoires qu'exerce la réalisation de certaines voyelles sur la réalisation des phonèmes consonantiques avoisinants. Par exemple, le /i/ de /bi/ demandera un

[23] Trad. « Bien que l'exemplaire de [ba] soit plus saillant que le [da], il y avait seulement deux options de réponses dans cette expérience (ba ou da), ce qui signifie que chaque différences dans leur saillance était minimisée (i.e., si l'indice pour un [ba] n'était pas présent, un [da] pouvait être présagé ».

étirement des lèvres qui contraindra la réalisation de la fermeture labiale. Un /a/ en revanche, qui est une voyelle ouverte, imposera mois de contraintes lors de sa réalisation, ce qui modifiera par conséquent moins la réalisation articulatoire de la consonne. Ainsi, les voyelles non arrondies améliorent l'intelligibilité des consonnes environnantes alors que les voyelles arrondies ou protruses diminuent au contraire cette intelligibilité.

Les constats faits pour les consonnes /b v ʁ l z ʒ/ s'étend aussi aux consonnes fricatives /s/ et /ʃ/. En effet, Benguérel & Pichora-Fuller (1982) montrent que ces dernières sont moins bien identifiées lorsqu'elles sont suivies d'une voyelle arrondie, ici /u/, que lorsqu'elles sont suivies d'une voyelle non arrondie comme /æ/. Notons que les voyelles qui contraignent le plus la réalisation des consonnes sont généralement les plus facilement identifiables visuellement. C'est le cas du /y/ dans l'expérience de Benoît et al. (1994) qui était la voyelle la plus facilement identifiable. Cependant, les variations de réalisation induites par la coarticulation ne gênent pas nécessairement l'identification. En effet, tout comme pour la perception auditive de la parole, les mouvements articulatoires peuvent être supprimés sans que cela ne modifie le taux d'identification. C'est ce qu'a montré la dernière expérience de Yakel (2000) en supprimant le noyau visuel[24] d'une voyelle insérée dans une séquence /bVb/. La suppression de cette information qui pourrait sembler indispensable pour identifier la voyelle n'impactait en rien les performances des participants, comme c'est le cas lors de la même modification du signal acoustique (Strange, Jenkins, & Johnson, 1983). Ce ne serait donc pas seulement le noyau de la voyelle ou même le *burst* (dans le cas des consonnes) qui serait détenteur de l'information sur l'identité du phonème. Des indices seraient également présents dans la coarticulation, et seraient transmis par les informations dynamiques (ceux-ci n'étant pas présents dans les paradigmes utilisant des protocoles statiques[25] (Summerfield & McGrath, 1984) qui montre séquentiellement des images). Des études ont montré que les modulation des gestes articulatoires de certains phonèmes sur d'autres pouvait être exploitée au cours des processus d'identification phonémique, en générant des « attentes coarticulatoires » (Mann, 1980). Dans ce sens et plus récemment, Fowler (2006) suggère que les interlocuteurs utilisent leurs connaissances implicites sur la production de la parole pour

[24]Le noyau visuel d'une voyelle ou d'une consonne est la partie stable de celle-ci qui contient les indices statiques qui permettent son identification

[25] Durant les protocoles de reconnaissance de voyelles (ou de consonnes) statiques, une configuration labiale fixe est présentée, celle-ci étant généralement la plus représentative du visème à identifier (par exemple, un locuteur la bouche ouverte permet d'identifier un /a/, des lèvres protruses un /u/ ou un /y/, etc).

guider sa perception. Cependant, d'autres hypothèses sont envisageables, notamment celle d'une association à long terme entre le signal et le percept (Lotto & Holt, 2006).

2.1.1.5.2 FACTEURS COGNITIFS, LANGAGIERS ET DIFFERENCES INTER-INDIVIDUELLES

Si bon nombre de facteurs ont été identifiés comme n'impactant pas l'utilisation des informations visuelles (voir Rönnberg, 1990 pour une revue). Par exmeple, le QI, le niveau éducatif, le genre, le statut économique, la présence et la durée d'une perte d'audition ne vont pas moduler l'exploitation des indices oro-faciaux dans les conversations face-à-face (Hygge, Rönnberg, Larsby, & Arlinger, 1992 ; Tye-Murray, Sommers, & Spehar, 2007a, 2007b). D'autres, en revanche, sont reconnus pour diminuer leur utilisation, qu'ils soient cognitifs, culturels ou développementaux. En termes de facteurs cognitifs, Andersson, Lyxell, Rönnberg et Spens (2001), ont mis en avant que la mémoire de travail et la rapidité de traitement (i.e., processus phonologique évalués par exemple lors de tâche de jugement de rimes) expliquaient 46% de la variance des performances.

Le poids donné aux informations visuelles varie entre les populations. Le pourcentage de réponses dites de « fusion » de type /da/ dans l'illusion McGurk, par exemple, diffère entre les anglophones et les japonais. Les japonais semblent moins influencés par les informations visuelles et produisent donc moins de réponses de fusion (Sekiyama & Tohkura, 1991). La première raison avancée pour expliquer la sous-utilisation de ces informations était liée au fait que les contacts visuels directs sont bien moins fréquents dans certains pays asiatiques que dans les pays d'Europe par exemple. Cependant, les résultats obtenus par Aloufy, Lapidot et Myslobodskym (1996) sur une population parlant l'hébreu remettent en cause cette interprétation et ont orienté les discussions vers l'impact d'un inventaire phonologique plus restreint. Lorsque ce dernier est composé de moins de phonèmes, il y a moins de chevauchement entre les catégories, ce qui facilite l'identification de chaque phonème sur une base auditive seule. Cela nécessite, de ce fait, moins d'aide des informations visuelles, les informations auditives étant à elles seules très informatives. Sekiyama & Burnham (2008) ont par la suite avancé l'idée que cet effet serait bien dû à une intégration moins importante chez les japonais. Chen & Massaro (2004) contrent cet argument, stipulant que le traitement de l'information est universel et que c'est bien la quantité d'information fournie par le canal visuel qui module l'utilisation de ces indices. Plus que le facteur culturel, se serait donc

l'organisation de l'inventaire phonologique ainsi que la saillance de ces unités qui guide l'utilisation des informations visuelles.

Enfin, il existe de nombreuses différences interindividuelles au niveau de nos aptitudes à décrypter le signal visuel de parole. A l'âge adulte, les différences de performance sont très marquées. Ainsi Demorest & Bernstein (1992) ont mené une étude de lecture labiale sur 104 sujets normo-entendants qui devaient reconnaître un maximum de mots dans des phrases isolées prononcées par un locuteur et une locutrice. Ils ont estimé que 4,9% de la variabilité des performances est due aux différences interlocuteurs et 10.5% de la variabilité était induite par les différences inter-participants (labiolecteurs expérimentés ou naïfs). Nous pouvons également ajouter que lors d'une présentation audiovisuelle, certains auditeurs se baseront préférentiellement sur l'une ou l'autre des modalités.

Ces facteurs expliquent pourquoi les capacités de lecture labiale en présentation visuelle seule peuvent varier de 5% d'identification correcte à 80% pour les participants les plus doués, avec des taux de réponses correctes parfois plus importants dans le cas de la présentation visuelle seule par rapport à la présentation audiovisuelle (Bernstein, Demorest, Coulter & O'Connell, 1991). Malgré ces différences, et ne considérant que les populations occidentales, nous faisons généralement tous assez bon usage de l'information visuelle. Si le signal acoustique est généralement suffisant pour comprendre notre interlocuteur, pourquoi utilisons-nous cette information ? Si, d'un œil naïf, les informations visuelles pourraient passer pour redondantes quand les informations auditives sont disponibles, ces deux sources sont en réalité complémentaires.

2.1.2 COMPLEMENTARITE DES SIGNAUX ACOUSTIQUE ET VISUEL

Le signal visuel de parole a la même origine que le signal auditif, ce dernier étant généré par les mouvements du conduit vocal lors de l'articulation. L'exploitation des mouvements articulatoires du locuteur est possible car ils sont en relation avec les mouvements du conduit vocal (Yehia, Rubin, & Vatikiotis-Bateson, 1998) et sont corrélés au signal acoustique (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Malgré cela, les deux flux nous apportent des informations qui, même si elles sont parfois redondantes, sont également complémentaires. En effet, les traits peu audibles sont généralement les plus visibles (Smeele & Sitting, 1991). Lors d'une présentation auditive de consonnes, les indices

de *mode d'articulation* sont les plus discriminants, alors que c'est la *place d'articulation* qui est le plus informatif lors d'une présentation des informations visuelles seules. Jesse & Massaro (2010) ajoutent également que les informations auditives nous renseignent sur le *voisement*, mais également la *place d'articulation* lorsqu'il est arrière. Les informations visuelles permettent également, dans une plus faible mesure, d'obtenir des renseignements sur le voisement (confirmant les observations de Benguerel & Pichora-Fuller (1982). Cependant, les auteurs constatent que l'association de l'information auditive et visuelle amène plus de réponses correctes que la somme des scores des deux modalités. Ce gain nous indique qu'il y aurait des indices supplémentaires lors de la présentation simultanée de ces deux sources. Robert-Ribes et al. (1998), qui testaient quant à eux des voyelles, ajoutent que lors d'une présentation auditive, le paramètre de *hauteur* est le plus robuste, suivi par *l'antériorité* et enfin *l'arrondissement*, alors que lors d'une présentation visuelle seule, *l'arrondissement* fournit plus d'informations que la *hauteur*, *l'antériorité* étant presque invisible. Tous ces paramètres sont mieux perçus lorsqu'ils bénéficient du double codage que lors de la présentation d'une seule modalité. L'avantage audiovisuel serait en effet dû, selon Campbell (2008), à la double implication des informations visuelles lors de la perception audiovisuelle. Celles-ci auraient à la fois un rôle de complétion du signal acoustique permettant entre autres de préciser le contenu du discours mais également un rôle de consolidation des informations auditives via la redondance de ces informations.

2.1.2.1 DECOURS TEMPOREL DU SIGNAL DE PAROLE AUDIOVISUELLE

Les deux rôles qu'attribue Campbell (2008) aux informations visuelles (i.e., consolidation et redondance) ne sont pas les uniques raisons de l'avantage audiovisuel. L'avantage audiovisuel peut être influencé par le fait que l'information articulatoire peut être disponible avant l'information acoustique.

2.1.2.1.1 « VISION LEADS AUDITION[26] »

Cette spécificité a été récemment étudiée par Chandrasekaran, Trubanova, Stillittano, Caplier et Ghazanfar (2009) qui s'intéressaient au décours temporel des deux types d'informations lors de la *production*. L'ensemble des résultats fourni par leurs études se base sur l'analyse de

[26] Trad. "La vision guide l'audition".

production de séquences tirées de corpus en anglais (enregistrements aux rayons-X ou cinéradiographie ou enregistrement audiovisuel) et français (enregistrement audiovisuel). A partir de ces signaux, des composantes visuelles (l'aire aux lèvres obtenue à partir des contours de lèvres) et auditives (*l'enveloppe*[27] *des « bandes étroites »* individuelles obtenue en appliquant une transformée de Hilbert sur les signaux auditifs filtrés en passe bande, et *l'enveloppe des « bandes larges »* obtenues par sommation des « bandes étroites ») ont été mesurés. A partir de ces mesures, les auteurs ont réalisé des corrélations entre la composante visuelle et les composantes auditives, des analyses spectrales des deux signaux afin d'estimer leur cohérence, ainsi que des mesures de « time-to-voice [28]». Leurs résultats montrent dans un premier temps une forte corrélation entre les contours des « bandes larges » et l'aire aux lèvres. Cela nous indique que les variations en termes d'aire aux lèvres sont temporellement similaires à celles de l'enveloppe auditive. Les moments durant lesquels l'ouverture labiale est importante coïncident avec les zones pour laquelle l'amplitude (unités de Hilbert) du signal acoustique est la plus importante. Cependant, leurs observations attestent également d'une importante variabilité inter- intra- individuelle puisque les corrélations obtenues sur la moyenne de toutes les phrases s'échelonnent de .31 à .65. Des correspondances entre les contours auditifs et l'aire aux lèvres pour des passages plus longs ont également été observés. De plus, les auteurs ont constaté que ces corrélations n'étaient pas dues à des variations arbitraires entre deux signaux étant donné que les analyses des corrélations aléatoires sont toujours significativement plus faibles que celles des signaux intacts. Les résultats pour le français sont similaires.

L'analyse spectrale a permis de montrer que les régions fréquentielles concernant le F1 (300-800Hz) et les F2 et F3 (3KHz) sont étroitement liées aux informations visuelles, ce qui pourrait être un indice exploitable lors de la présentation de voyelles dans le bruit. Le signal visuel pourrait contenir des indices quant aux bandes de fréquence produites. Enfin, le délai de « time-to-voice » a également été analysé pour les consonnes /p b m/ et /f/. Ce délai était de 195 ms plus ou moins 60 ms pour le /p/, de 137 ms plus ou moins 64 ms pour le /m/, 205 plus ou moins 69 ms pour le /b/ et 240 plus ou moins 38 ms pour le /f/. Les délais « time-to-voice » apparaissent comme étant raccourcis lorsqu'ils ne se situent pas à l'attaque d'un mot. Cette étude permet donc d'une part de mettre en évidence que la réalisation visuelle (i.e.,

[27]Amplitude instantanée.

[28]Délai entre le début des mouvements labiaux et le début de la vocalisation. Celui-ci est déterminé par le temps entre l'onset du signal d'intérêt (e.g., *burst* acoustique) et la moitié de la trajectoire de fermeture labiale (déterminée à partir du début de la fermeture labiale).

aire aux lèvres) et la réalisation acoustique (i.e., enveloppe des bandes larges) sont largement corrélées, mais également d'attester, d'une avance temporelle importante des informations visuelles sur les informations auditives (cf. Chapitre 2.1.2.1.2 « Audition leads vision » ci-dessous pour une contre-argumentation).

Quelques études ont été menées sur le décours temporel de la parole audiovisuelle lors de la *perception.* Celles-ci avaient pour but de vérifier si l'observation des indices fournis par les gestes oro-faciaux, notamment leur avance sur le signal acoustique, pouvait entraîner un bénéfice en termes d'identification. Cathiard (1994) a étudié l'impact de l'anticipation labiale lors de pauses silencieuses. Pour cela, elle a testé la transition vocalique /i → y/ (par rapport à la transition contrôle /i → i/). Le but de son étude était de voir si durant la pause, les mouvements anticipatoires de protrusion pour réaliser le /y/ (passage d'une articulation étirée à arrondie) étaient utilisés par les participants pour déterminer l'identité de la voyelle qui suivait. Ces transitions étaient insérées dans une phrase porteuse de type « Tu dis /y/ ? » ou « Tu dis /i/ ?». La pause entre le /i/ et le /y/ ou entre le /i/ et le /i/ était soit longue (460 ms) soit courte (160ms). Les participants devaient déterminer l'identité de la voyelle qui allait apparaitre après la pause dans deux conditions : auditive et audiovisuelle. Les résultats en modalité audiovisuelle montrent une anticipation de la voyelle /y/ (basée sur les informations visuelles). En effet, les sujets identifiaient cette voyelle durant les pauses, avant même que l'information auditive ne soit disponible. En réalité, les sujets se basaient sur le trait d'arrondissement du /y/ qui était décelable bien avant l'information acoustique de la voyelle. Cette anticipation était de l'ordre de 60 ms pour la petite pause et 160 ms pour la grande pause. Ainsi, le décours temporel de l'anticipation dépendrait de la longueur de la pause. Cependant, en modalité auditive, l'identification de la voyelle n'était possible qu'après l'apparition de l'information acoustique de la voyelle. Ainsi, nous observons bien une précédence de l'information visuelle sur le signal acoustique qui est utilisée pour identifier la voyelle. Signalons que Escudier, Benoît et Lallouache (1990) avaient quelque années auparavant tiré les mêmes conclusions en utilisant la même transition (i.e., 40-60 ms d'anticipation du signal visuel sur le début acoustique de la voyelle) avec des productions naturelles de séquences de type « /zizy/ ».

La plupart des expériences qui ont étudié le décours temporel des informations auditives et visuelles en perception de la parole ont utilisé un paradigme de dévoilement progressif ou *gating.* Dans le cadre de la parole auditive, il s'agit de présenter le signal

acoustique en séquence (*gates*) de longueur croissante afin de ne dévoiler, dans les premiers essais, que les premières informations contenues dans le signal jusqu'à leur intégralité. Pour la parole audiovisuelle, le signal auditif, mais également le signal visuel, peuvent être dévoilés de manière progressive. Ce paradigme permet d'étudier l'évolution de l'intégration des deux modalités et d'étudier comment chacune des modalités contribue à l'avantage audiovisuel (Grosjean, 1980, 1996 ; Jesse & Massaro, 2010 ; Moradi, Lidestam, & Rönnberg, 2013 ; Munhall & Tohkura, 1998 ; Troille, Cathiard, & Abry, 2010). C'est notamment le cas de l'étude de Smeele (1994), qui s'est s'intéressée à l'identification audiovisuelle des consonnes /b, d, p, t, k, v, f, z, s, m, n/. Des séquences CV dont la voyelle était toujours /a/ étaient présentées sous forme de *gates* (cinq par syllabe) de 40 ms en modalité auditive-seule, visuelle-seule et bimodale. Cette étude a pu ainsi mettre en évidence que les bilabiales et labiodentales étaient identifiées de manière plus précoce quand le signal visuel était disponible. Le fait de voir la fermeture labiale du locuteur permet d'identifier la place d'articulation avant même que les informations auditives ne soient disponibles.

Munhall & Tohkura (1998) ont voulu étudier le décours temporel des informations auditives et visuelles lors de la perception de consonnes utilisant la tâche de *gating* avec l'effet McGurk. Pour ce faire, ils ont présenté l'intégralité du signal auditif /æ**b**æ/ à chaque essai (ce signal n'était donc pas présenté en « *gating* ») alors que les informations visuelles d'un /æ**g**æ/ étaient dévoilées progressivement au fur et à mesure des essais. Dans la seconde partie de l'expérience, c'est le signal visuel qui était présenté entièrement alors que le signal auditif était présenté en segments. Les résultats ont révélé que le nombre de réponses /d/ augmentait de manière linéaire au fur et à mesure que l'information visuelle était dévoilée. A l'inverse, quand le signal auditif était présenté sous forme de *gates*, les réponses de fusion n'apparaissaient qu'à partir du moment où le *burst* était présenté ; le taux de réponse « fusion » n'augmentait plus par la suite. Ils ont donc montré que les informations visuelles et auditives ne contribuent pas de la même façon à l'illusion car celles-ci ne se dévoilent pas au même rythme. Cependant, les auteurs n'avaient considéré qu'un nombre réduit de consonnes et notamment uniquement le /b/ auditif, bénéficiant d'un indice temporellement délimité sur l'identité de la consonne (comparé à un indice de friction qui lui est distribué sur une partie

plus importante du signal). D'autres phonèmes ont des indices plus distribués[29] sur le plan auditif quant à leur identité (Smits, Warner, McQueen, & Cutler, 2003 ; Smits, 2000).

Jesse & Massaro (2010) ont étudié cette problématique surtout dans le cadre de la reconnaissance des mots en condition multimodale. Avec de la parole (auditive et visuelle) synthétique et un paradigme de *gating*, ils ont observé le décours temporel des informations de chaque modalité. Leurs stimuli étaient composés des 22 consonnes initiales de l'anglais insérées dans des séquences CVC précédées de la voyelle /a/ (que les participants ne devaient pas considérer) de type « a cash ». Les séquences étaient divisées en 6 *gates*. Leur durée n'était pas fixe mais proportionnelle. La durée des trois premiers *gates* équivalait à un tiers de la durée de la première consonne (*M* = 35ms, rang = 12-64 ms). Ainsi, à la fin du troisième *gate*, toute la première consonne était présentée. Le premier *gate* contenait toujours les mouvements préparatoires de la coarticulation. Les auteurs ont observé une amélioration des performances pour tous les *gates* et tous les phonèmes en condition audiovisuelle par rapport à la condition auditive seule. Leurs résultats ont montré que les informations visuelles (notamment les indices relatifs à la place d'articulation) sont présentes et utilisées de manière précoce dans le signal (i.e., durant les deux premiers *gates*[30]). Ils ont également montré que leur utilisation augmente peu au cours des *gates* suivants alors que les informations auditives sur l'identité du phonème s'accumulent au cours du temps. Ils notent que ces résultats sont conformes à ceux présentés par de la Vaux et Massaro (2004). Ils contrastent partiellement avec les observations de Munhall et Tokura. Ces derniers avaient observé que les informations transmises par le canal acoustique variaient rapidement et de manière non linéaire. Ils ne testaient cependant que trois consonnes plosives ce qui peut expliquer leur pattern de résultats différent par rapport à ceux de Smits, Warner, McQueen et Cutler (2003) et Smits (2000), qui avaient obtenu des patterns linéaires pour un large panel de consonnes présentées sous forme de *gating* auditif. Dans des séquences où une voyelle est prononcée avant la consonne d'intérêt, comme cela était le cas dans l'étude de Jesse & Massaro (2010), il s'avère que les informations auditives sont plus informatives que les informations visuelles. Toutefois, comme l'écrivent les auteurs, « *the available visual information already occurs close to its maximal level of transmission early in the consonant, whereas the auditory information is more distributed across the phoneme. Visual speech information thus*

[29]Le signal de friction est distribué dans le sens où celui-ci s'étend dans le temps. Un friction dans un mot peut durer plusieurs centaines de millisecondes alors qu'un *burst* acoustique est ponctuel.

[30]Contenant les mouvements préparatoires ainsi que le début articulatoire voire acoustique de la consonne.

supplements early-on auditory information and leads to robust early recognition benefits, despite acoustic information that is already available[31] » (p.222). Les auteurs ont observé un bénéfice audiovisuel précoce pour le place d'articulation (comme chez Smeele, 1994) mais également pour l'arrondissement, la friction (qui n'était alors pas obtenu par Smeele, 1994) et la durée. Lorsque tout le phonème était présenté, un avantage audiovisuel a été obtenu pour la nasalité et le voisement en plus des caractéristiques précédemment citées.

Notons que les stimuli de Jesse & Massaro (2010) étaient élaborés à partir de signaux de parole synthétique, ce qui ne limite pas leurs conclusions, étant donné que celles-ci sont cohérentes avec les résultats précédemment obtenus dans la littérature. Néanmoins, il est important de préciser que des stimuli synthétiques ne contiendront jamais toutes les variations de la parole naturelle, qu'elle soit auditive ou visuelle. Comme le soulignent Abel et al. (2011), des détails visuels fins peuvent être utilisés pour discriminer des phonèmes qui appartiennent à la même catégorie visémique (e.g., /b p m/).

La prise en compte des indices auditifs et visuels de la parole semble être très robuste car les propriétés intrinsèques à la parole multimodale semblent favoriser et faciliter la perception des phonèmes. Même lorsque le décours temporel ne respecte pas celui de la parole naturelle, c'est-à-dire dans les cas d'asynchronie son-image, l'intégration bimodale a quand même lieu. Dans une expérience d'identification de syllabes, Smeele (1994) a observé que le bénéfice de l'information visuelle était toujours présent sur des plages de désynchronisation son-image relativement importantes. Cependant, quand le signal auditif arrivait avant signal visuel, la performance était plus détériorée que lorsque c'était le signal visuel qui survenait en premier. Les deux sources d'informations pouvaient encore être intégrées avec une asynchronie de 300 ms. Abry, Lallouache et Cathiard (1996) ainsi que Cathiard & Tibergihien (1994 cité par Colin & Radeau, 2003) signalent que le retard du son pouvait être comblé dans la mesure où le début du son n'apparaît pas après la fin de l'articulation visible.

Pour résumer, il semblerait qu'en production les informations visuelles précédent, dans certains cas, les informations auditives, notamment dans les séquences à consonnes initiales, et fournissent des indices qui sont corrélées au signal auditif qui suit (e.g., les variations

[31] Trad. "Les informations visuelles disponible sont déjà proche de leur niveau maximum de transmission précocement dans la consonnes, alors que les informations auditives sont plus distribuées tout au long du phonème. Les informations visuelles de parole complètent ainsi de manière précoce et conduit à de solides avantages de reconnaissances, malgré le fait que les informations acoustiques sont déjà disponible".

temporelles de l'aire aux lèvres sont similaires aux variations de l'enveloppe auditive et celles-ci renseignent notamment sur les bandes de fréquences du signal auditif ; Chandrasekaran et al., 2009). En perception, ces indices fournis par l'articulation visible sont utilisés pour anticiper l'identité des phonèmes prononcés. L'information visuelle est utilisée de manière précoce et renseigne notamment sur la place d'articulation. Cette anticipation est d'autant plus importante que l'articulation est visible (Cathiard, 1994 ; Jesse & Massaro, 2010 ; Smeele, 1994 ; de la Vaux et Massaro, 2004). L'utilisation des indices fournis par le canal visuel est également modulée en fonction de la quantité de signal dévoilé ainsi que de la structure de la séquence (i.e., CV ou VC) (Jesse & Massaro, 2010). Car en effet, alors que les informations visuelles précédent parfois les informations auditives (notamment dans les séquences CV), le phénomène inverse a également été observé, où l'information auditive peut être utilisée avant l'apparition de l'information visuelle.

2.1.2.1.2 « AUDITION LEADS VISION[32] »

La précédence des informations visuelles sur l'information auditive, autrement dit le fait que « vision leads audition » n'est pas un phénomène constant. En effet, il semblerait que le signal auditif soit parfois porteur d'informations qui ne seront disponibles que plus tard dans le signal visuel. Troille, Cathiard et Abry (2007, 2010) montrent que l'identité d'une voyelle peut être détectée sur la base des informations auditives 40 à 60 ms avant que l'information visuelle ne soit traitée. Leurs résultats obtenus en 2010 montrent en effet que les participants peuvent détecter le /y/ dans une séquence /zizy/ 92 ms avant la fin de la friction. En modalité audiovisuelle cependant, la voyelle arrondie n'est détectée que 73 ms avant la fin de la friction (Figure 10).

[32] Trad. « L'audition guide la vision ».

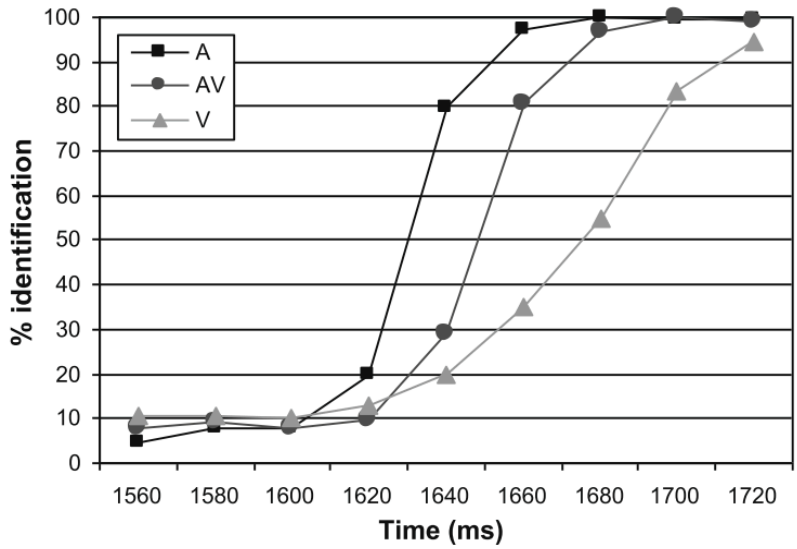

Figure 8. Pourcentage de détection correcte du /y/ dans des séquences /zizy/ en fonction de la modalité de présentation. Le bruit de friction du second /z/ se trouve entre 1600 et 1720ms. (Tiré de Troille et al., 2010)

Plus récemment, Schwartz & Savariaux (2013) ont également mis en évidence que les relations temporelles entre les indices visuels et auditifs sont très complexes. Il semblerait que lors de la *production*, les informations visuelles précèdent l'information acoustique pour les phonèmes à l'initiale d'une séquence. Dans ce cas, les auteurs ont observé que les informations visuelles de fermeture labiale pouvaient avoir plus de 350 ms d'avance sur les premiers indices acoustiques de fermeture (i.e., diminution de l'intensité et diminution de la F1). Concernant les séquences /aCa/, lors de la fermeture, une légère avance des informations visuelles est observée sur les informations acoustiques, de l'ordre de 20 à 80 ms (en faveur des informations sur l'intensité). Enfin, la synchronie entre les deux canaux est presque parfaite pour des bilabiales de milieux de séquence lors de l'ouverture labiale (voir aussi Schwartz & Savariaux, 2014). Cependant, le gain lié aux informations visuelles n'est pas remis en question. Selon les auteurs, même lorsque les entrées visuelles ne sont pas présentes en amont du signal acoustique, elles permettent tout de même un gain en termes de prédictibilité. Le fait que la vision guide l'audition permet d'ailleurs un gain temporel plus important (170 ms) que dans les cas où les informations auditives peuvent être utilisées avant les information visuelles (30 ms) (Massaro, Cohen, & Smeele, 1996 ; Munhall, Gribble, Sacco, & Ward, 1996 ; van Wassenhove, Grant, & Poeppel, 2007).

2.1.2.1.3 LE POTENTIEL PREDICTIF DE LA PAROLE AUDIOVISUELLE

2.1.2.1.3.1 LE « QUOI » ET LE « QUAND »

Ces études laissent penser que le bénéfice audiovisuel dépend de la distribution différentielle des informations fournies par les deux modalités au cours du temps. Ces différences, et notamment l'avantage temporel pour la reconnaissance fourni par la présentation audiovisuelle, permettraient de prédire le contenu du signal acoustique qui va suivre. Ce potentiel prédictif de la parole visible n'a été que récemment montré. La présence d'information visuelle pertinente et exploitable permettrait à l'auditeur de prédire le contenu acoustique et engendrerait un effet de facilitation des processus perceptifs de parole. Le système tirerait partie de la précédence des informations visuelles en permettant une anticipation, engendrant par là même des prédictions sur le signal acoustique qui va suivre. Cela produit notamment une accélération des réponses corticales au niveaux des aires auditives (Arnal, Morillon, Kell, & Giraud, 2009 ; Besle, Fort, Delpuech, & Giard, 2004 ; Stekelenburg & Vroomen, 2007 ; Wassenhove, Grant, & Poeppel, 2005). Il reste à comprendre comment les informations visuelles permettent cet avantage.

La littérature a pris deux chemins différents, mais pas opposés, pour répondre à cette question. L'information visuelle donnerait des indices à la fois sur le « quoi » (i.e., quelle est l'identité du phonème qui va suivre) en agissant donc comme une amorce, et sur le « quand » (i.e., quand le signal acoustique va débuter). En 2013, Paris, Kim et Davis ont montré au travers de deux expériences que la saillance visuelle entrainait une facilitation des processus perceptifs (observée sur les temps de réponse) ; celle-ci était davantage due à des indices de forme qu'à des indices temporels. Ce n'est donc pas une prédiction sur le « quand » l'information auditive va arriver, mais sur le « quoi », qui serait véhiculée par les informations visuelles. Cependant, ces auteurs n'ont pas obtenu le classique effet de facilitation plus important pour les consonnes plus visibles, comme c'est le cas dans les études en potentiels évoqués de référence (Arnal et al., 2009 ; Stekelenburg & Vroomen, 2007 ; van Wassenhove et al., 2005). Des critiques méthodologiques ont été formulées par les auteurs eux-mêmes car une tâche à choix forcé peut réduire la saillance perceptive des séquences (car en effet, si les indices indiquant un /ba/ ne sont pas présents, le /da/ sera choisi). Le type de tâche qu'ils ont utilisé (i.e., identification) explique également l'absence d'effet des indices temporels (indice de « quand »). En effet, la tâche d'identification rend saillante les informations de forme. Cela peut expliquer pourquoi leurs observations, sur ce point, divergent des résultats obtenus par Schwartz et al., (2004). Ces derniers obtenaient une facilitation qui serait due aux indices temporels (i.e., « quand »). Ils observaient en effet que les mouvements de la langue pour

articuler le /u/ ou /y/ étaient initiés 100 ms avant le dévoisement et 240 ms avant le noyau de la voyelle ce qui fournit des indices temporels (et non d'identité puisque les deux voyelles sont très proches visuellement) ; ceci faciliterait la détection de la cible à venir. Donc même si les indices de forme permettent un avantage temporel, les indices temporels restent pertinents dans le cadre de scores d'identification.

Ces études mettent en évidence que les *indices de forme* permettent un effet d'amorçage qui réduit le coût des traitements subséquents (notamment l'activité du cortex auditif comparé à une condition sans indice) en permettant de prédire quelles caractéristiques acoustiques vont suivre (Jääskeläinen et al., 2008 ; Paris et al., 2013) et dans l'autre cas que des *indices temporels* permettent une amélioration des scores d'identification (Schwartz et al., 2004). Certains travaux ont également suggéré que c'est tout particulièrement l'aire aux lèvres qui semble être codée; précisant plus tard que d'autres paramètres (e.g., l'étendue de la protrusion) entrent aussi en ligne de compte (Cathiard, 1994 ; Abry & Lallouache, 1995 ; Lallouache, 1991).

L'exploitation des mouvements articulatoires visibles du locuteur fournit donc un avantage, d'une part car cette information est complémentaire au signal acoustique, mais également car elle est souvent présente avant l'information acoustique, permettant donc des prédictions aussi bien sur la forme de ce qui va suivre que sur la temporalité des informations subséquentes. Au niveau neuronal, il a également été montré que les informations visuelles modulent et facilitent les traitements réalisés par le cortex auditif.

2.1.2.2 ETUDES NEUROPHYSIOLOGIQUES

2.1.2.2.1 IMPACT DE LA MODALITE VISUELLE SUR LE CORTEX AUDITIF

De nombreuses recherches ont mis en évidence que la parole visuelle influence l'activité du cortex auditif primaire. Alors que nous savons que les sons simples sont traités de manière préférentielle dans le cortex auditif primaire, de nombreuses études ont montré que plus les entrées étaient porteuses de sens, plus le traitement était relayé par le STS, STG ou le gyrus temporal médian, voire des aires de plus haut niveau comme les régions motrices ou fronto-pariétales (Hickok & Poeppel, 2000 ; Gregory Hickok & Poeppel, 2007; Pasley et al., 2012 ; Patterson, Uppenkamp, Johnsrude, & Griffiths, 2002). Certains pensent que la vision intervient dans ces traitements, notamment pour renforcer la repondération des informations du système auditif, permettant la fusion des représentations auditives plus complexes avec les

représentations visuelles. En effet, cela semble un bon moyen, dans des conditions d'écoute naturelle qui sont souvent bruitées, d'augmenter l'intelligibilité afin de tirer le meilleur parti des signaux d'entrée.

Sams et al. (1991) furent parmi les premiers à montrer, via la technique de *mismatch negativity* (MMN) en MEG, que les informations visuelles liées à la parole modulaient l'activité du cortex auditif. Il demanda à dix adultes de compter le nombre de syllabes qui leur étaient présentées de manière audiovisuelle, celles-ci étant un /pa/ et un /ka/ dont l'information visuelle pouvait être congruente aux informations auditives (V=A) ou incongruentes (/pa/ auditif et /ka/ visuel : V≠A). Les stimuli incongruents généraient toujours les percepts /ta/ ou /ka/ chez les participants. Le ratio des stimuli présentés était également manipulé avec 84% de V=A et 16% de V≠A ou inversement. Lorsque les stimuli V=A étaient fréquents, trois ondes consécutives étaient observées. Cependant, une modification des potentiels évoqués après 180 ms (durée à partir de laquelle les deux stimuli étaient différents) a été observée pour les stimuli V≠A peu fréquents (les potentiels générés à 50 et 100 ms étant similaires dans les deux conditions). La même observation fut faite pour les stimuli V=A peu fréquents. Ce n'est donc pas la congruence, ou la nature des informations visuelles qui produiraient cet effet, car seuls les stimuli peu fréquents induisaient une modulation de la réponse supra-temporale (i.e., l'effet a été observé lors de la perception de /ta/ ou /ka/ résultant d'une illusion (V≠A) présentés parmi beaucoup de /pa/ (V=A) ou lors de la perception de /pa/ (V=A) présentés parmi beaucoup de /ta/ ou /ka/ résultant d'une illusion (V≠A)). Cela signifie donc qu'une réponse de *mismatch* est provoquée par l'évènement /pa/ peu fréquent, présenté parmi beaucoup de /pa/ auditifs accompagné de la production visuelle d'un /ka/. Alors que les entrées auditives sont les mêmes, c'est la résultante de l'illusion audiovisuelle qui produit une MMN lors de la perception d'un /pa/ audiovisuel. En d'autres termes, une MMN est évoquée par un phonème illusoire induit par une stimulation audiovisuelle. La distribution des différences observées serait selon les auteurs la résultante de l'impact des informations visuelles sur le cortex auditif.

Colin, Radeau, Soquet, Dachy et Deltenre (2002) ont, avec le même paradigme, mais en utilisant le ventriloquisme, montré que ces illusions se mettent en place précocement lors du traitement perceptif. Par la suite, Möttönen, Krause, Tiippana et Sams (2002) ont également retrouvé des résultats similaires. Il semblerait donc que la composante visuelle entre en interaction avec les informations auditives entre 150 et 250 ms. Les résultats obtenus

en EEG et MEG ont été confortés par des études de neuro-imagerie fonctionnelle (Calvert, 2001). Plus récemment, Okada, Venezia, Matchin, Saberi et Hickok ont mené en 2013 une étude IRMf dont le but était de savoir comment la parole visuelle pouvait influencer l'activité du cortex auditif avant et après sa réponse à de la parole auditive. Pour ce faire, ils ont présenté aux participants les syllabes /ra/, /la/, /ma/ et /na/ en situation audiovisuelle et auditive seule. Ils ont analysé une région d'intérêt (ROI : *Region Of Interest*), le cortex auditif, déterminée par le contraste entre présentation avec bruit > repos. Cette analyse a montré une activité plus importante lors de la présentation audiovisuelle par rapport à la présentation auditive seule. L'ajout de l'information visuelle module donc bien l'activité du cortex auditif primaire.

2.1.2.2.2 INTERACTION SUPPRESSION

L'impact de la présentation bimodale sur les traitements du cortex auditif est double. Il permet de réduire le coût des traitements auditifs (Besle, Fort, Delpuech, & Giard, 2004 ; van Wassenhove et al., 2005), mais également d'accélérer ces traitements (Arnal et al., 2009 ; van Wassenhove et al., 2005). De nombreuses études en EEG ont été réalisées afin de comprendre comment les informations visuelles interagissaient avec le cortex auditif. Celles-ci ont permis de mettre en évidence que ces interactions sont plus précoces que celles observées dans les études antérieures. Elles ont utilisé des potentiels évoqués (ERP : Event Related Potentials) car ils permettent de mesurer l'évolution temporelle des processus perceptifs. Les potentiels évoqués sont des variations de l'activité électrique cérébrale en réponse à un stimulus donné. Ils sont extraits de l'électroencéphalogramme en moyennant les signaux produits par cette stimulation. Les potentiels évoqués sont généralement identifiés par leur polarité (positive ou négative), la latence (temps de déclenchement de la réponse par rapport au début de la stimulation, mesuré en millisecondes), l'amplitude (mesuré en μV) ainsi que la distribution topographique des modulations électrophysiologiques à travers le scalp (Coles & Rugg, 1995). L'amplitude reflète l'intensité des réponses neuronales, alors que la latence fait référence au moment où l'on observe un pic d'amplitude produit par la réponse au stimulus. Les pics précoces fournissent donc des informations sur les processus qui sont réalisés au tout début du processus alors que les pics plus tardifs rendent compte des mécanismes opérant à des étapes subséquentes du traitement perceptif.

Quelques travaux sur la perception de la parole audiovisuelle ont étudié les réponses électrophysiologiques produites par des potentiels évoqués auditifs précoces, c'est-à-dire, des processus auditifs ayant lieu dans les toutes premières millisecondes du traitement de l'information visuelle et audiovisuelle. Les potentiels auditifs précoces consistent en une séquence de pics connus comme le complexe N1/P2. Ce complexe est produit par des stimuli auditifs pour lesquels on observe un pic négatif autour de 100 à 150 ms (N1) et un pic positif dans les 200 à 250 ms (P2) après le début de la stimulation. Le rôle fonctionnel de N1 est en rapport avec les processus permettant la détection et l'encodage des propriétés auditives de la stimulation. Sa source neuronale a été localisée dans le cortex auditif (Eggermont & Ponton, 2002 ; Näätänen & Picton, 1987). Le rôle de P2 est moins clair. Il est souvent décrit comme une période de positivité après l'occurrence de N1 et faisant partie du pattern d'activations. Il est possible que l'amplitude de P2 soit modulée par des facteurs contextuels, comme par exemple l'attention.

Des études ERP sur la perception de la parole audiovisuelle indiquent que la détection visuelle des mouvements de préparation articulatoire ajoutée à celle de la détection auditive accélère le déroulement temporel des traitements acoustiques (Stekelenburg & Vroomen, 2007). En particulier, l'étude de van Wassenhove et al. (2005) a mis en évidence que le traitement visuel des gestes labiaux « préparateurs » en présentation audiovisuelle (e.g., les mouvements labiaux qui précèdent la fermeture labiale et qui conduisent à la production acoustique du phonème /p/ dans la syllabe /pa/) diminue les latences des composantes évoquées N1 et P2 au niveau des aires auditives. Lors d'une présentation visuelle seule, ils ont observé des pics à environ 400 ms avant le début acoustique du stimulus qui n'a pas donné lieu à un potentiel évoqué auditif ; ils ont observé des potentiels évoqués visuels typiques dans les zones temporo-occipitales. Enfin, il apparaît que la magnitude du décalage dans la latence du complexe N1/P2 dépend de la saillance visuelle du geste de production. Par exemple, /p/, qui est produit avec un geste labial, est plus saillant visuellement que /k/ (comme dans /ka/), qui est produit grâce à une occlusion de la partie postérieure de la cavité buccale. Les résultats des ERP montrent que /p/ produit des décalages de latence en N1 et P2 plus importants que /k/. Ce phénomène est nommé « interaction suppression » du cortex auditif.

Cette suppression de la N1 auditive ou désactivation corticale serait liée à une facilitation du traitement des indices acoustiques, qui serait par conséquent plus rapide (Besle et al., 2004). Selon les auteurs, elle serait due au fait que les informations labiales, qui

précèdent les informations auditives (à cause de la coarticulation anticipatoire), réduirait l'incertitude du signal et diminuerait les demandes computationnelles des aires du cortex auditif (Besle et al., 2004 ; Wassenhove et al., 2005). Ces « désactivations » corticales dues à la bimodalité refléteraient une facilitation des traitements auditifs. Cette dernière serait associée à la facilitation comportementale qui accélère l'identification de la syllabe présentée en situation audiovisuelle par rapport à une présentation auditive seule (Besle et al., 2004 ; Klucharev, Möttönen, & Sams, 2003). Cependant, et comme l'on montré Stekelenburg & Vroomen (2007), il semblerait que cette suppression ne soit pas spécifique à la parole, car celle-ci intervient également pour des évènements écologiquement valides comme taper sur une tasse avec une cuillère. Il est également à noter que cette modulation de N1 dans le cortex auditif n'est pas influencée par la non-congruence des informations auditives et visuelles (e.g., /fu/ auditif couplé à un /bi/ visuel ou un claquement de main auditif couplé à la vidéo d'une cuillère qui tape sur une tasse). Enfin, cette suppression de N1 n'apparaît que lorsque des mouvements anticipatoires sont accessibles, comme le montre le résultat de leur expérience n°3 dans laquelle ils utilisaient des vidéos montrant deux mains déchirant une feuille de papier, vidéo dans laquelle aucun mouvement anticipatoire n'était disponible. L'apport principal de cette étude est de mettre en évidence que l'intégration audiovisuelle n'est pas spécifique à la parole, contrairement à ce qui avait été précédemment avancé.

Certains auteurs n'ont pas réussi à trouver une suppression de N1 lors de la présentation de stimuli mettant en jeu des formes géométriques simples et des tons purs (Fort, Delpuech, Pernier, & Giard, 2002 ; Giard & Peronnet, 1999). Stekelenburg & Vroomen (2007) avancent que c'est le sens de la relation entre les deux types d'informations, notamment le fait que les informations visuelles fournissent des informations pertinentes, si ce n'est sur le contenu phonologique, au moins sur le fait qu'un entrée acoustique va succéder à ces mouvements préparatoires, qui permettent cette intégration. De plus, cette étude nous informe sur la dissociation qui existe entre N1 et P2. En effet, la N1 n'est pas modulée par le caractère congruent des stimuli, elle dépend principalement du fait que le signal visuel contient ou non des mouvements anticipatoires. Lorsqu'il n'y a pas de mouvements anticipatoires, l'effet intermodal sur la N1 disparaît. Ce serait donc bel et bien la relation temporelle entre les informations auditives et visuelles qui serait importante pour l'intégration audiovisuelle, et pas le contenu du son à proprement parler. A l'inverse, la P2 est dépendante du contenu, celle-ci subissant une réduction plus importante lorsque le contenu des stimuli est incongruent.

En étudiant seulement la N1, Bhat, Pitt et Shahin (2014) ont voulu tester l'hypothèse proposée par Chandrasekaran & Ghazanfar (2009) et par Luo, Liu et Poeppel (2010) selon laquelle l'intégration audiovisuelle serait liée à l'alignement temporel de la réponse neuronale aux mouvements de la bouche avec la réponse représentant le contour de l'enveloppe sonore de la parole[33]. Les modulations temporelles lentes du signal acoustique de parole qui vont de 0 à 10 Hz en fonction des études (2-7 Hz pour Chandrasekaran & Ghazanfar, 2009 ; 0-10 Hz pour Munhall & Vatikiotis-Bateson, 1998 ou 5 Hz pour Ohala, 1975) seraient la clé de l'intégration audiovisuelle. Ces fréquences sont d'ailleurs cruciales pour la reconnaissance de la parole, aussi bien au niveau auditif (Drullman, Festen, & Plomp, 1994 ; Drullman, 1995) que visuel. Par exemple, dans le cadre des informations visuelles, la réduction (Rosenblum, 2005) ou l'augmentation (Kim & Davis, 2004) des fréquences des images lors de la perception audiovisuelle supprime l'avantage audiovisuel. Bhat et al. (2014) ont utilisé l'effet de restauration illusoire pour tester l'hypothèse précédemment décrite. L'idée est que s'il y a bien une cohérence temporelle entre les informations visuelles et le contour de l'enveloppe acoustique, et que ce lien est critique pour l'intégration, alors le fait de bruiter une partie du signal acoustique ne devrait pas empêcher l'intégration car le système pourrait se baser sur l'information visuelle pour « récupérer » cette information. Les auteurs suggèrent que lors de la perception de la partie bruitée du signal, une désactivation du cortex auditif induite par le cortex visuel au moment de cette interruption devrait être observée. Cela devrait se manifester par un nombre plus important de réponse « illusion » (i.e., le participant ne perçoit pas d'interruption dans le signal acoustique alors qu'il y en a une) au niveau comportemental lors de la présentation de stimuli congruents par rapport aux stimuli incongruents (i.e., où les mouvements articulatoires sont temporellement inversés). De plus, la désactivation du cortex auditif devrait être plus importante dans le cas où les stimuli congruents ont entrainé une illusion chez les participants. Bhat et al. (2014) ont en effet montré que lorsqu'un segment acoustique est bruité, mais que les informations visuelles sont disponibles, il y a un effet de suppression des réponses dans le cortex auditif au début et à la fin de l'apparition de la zone bruitée. Cette suppression des réponses auditives est induite par le cortex visuel, créant

[33] L'enveloppe temporelle de la parole, que nous avons déjà évoquée au Chapitre 3.1.2.1.1 "Vision leads Audition », correspond à l'ensemble des modulations d'amplitude de basse fréquence, généralement de 2 à 50 Hz, qui correspondent aux transitions syllabiques ou phonétiques. Elles sont critiques pour la compréhension de la parole (Abramson & Lisker, 1970), notamment la segmentation. La parole peut en effet être reconnue lorsque les informations spectrales sont très limitées mais que les indices de l'enveloppe temporelle sont conservés (Fisher, 1968).

l'illusion que l'enveloppe sonore est continue. Cet effet est présent uniquement pour les stimuli congruents. Le contexte visuel augmente donc le processus d'inhibition, ce qui renforcerait d'autant plus l'illusion perceptive. Cette hypothèse semble en accord avec les résultats de Golumbic, 2013 qui montrent que la présentation de parole auditive en présence d'informations labiale augmente l'encodage de l'enveloppe de la parole dans le cortex auditif lorsque les informations auditives ne sont pas en elles mêmes suffisantes pour comprendre le discours, comme lors du *Coktail Party Effect*.

Une autre interprétation possible pour cette interaction audiovisuelle serait que les mouvements visuels entraîneraient un « filtrage » sensoriel des informations auditives. Ce « filtrage sensoriel » est décrit comme « *the ability of the brain to modulate its sensitivity to incoming stimuli*[34] » (Kim & Davis, 2004, p. 917). Cela inclue le fait de filtrer (réduire = *gate out*) les entrées non pertinentes ou de ne laisser passer (*gate in*) que les entrées sensorielles nouvelles ou qui font sens. Cela entraînerait des réponses similaires à celles des paradigmes de répétition suppression. En effet, lorsque l'on présente en modalité auditive un même son à 1 sec et à 500 ms d'intervalle, on observe alors une suppression des potentiels auditifs (P50, N1, P2) qui s'accompagne d'une réduction de la latence de N1 (Johannesen et al., 2005 ; Kizkin, Karlidag, Ozcan, & Ozisik, 2006).

L'information visuelle jouerait le même rôle, en fournissant des informations redondantes qui entraîneraient la suppression des potentiels auditifs. Cela est observé lors de présentation bimodale de clics précédés d'un flash lumineux (Oray, Lu, & Dawson, 2002) et serait interprétable comme un corrélat neuronal d'un filtrage sensoriel intermodal. Cela est corroboré par les résultats de Stekelenburg & Vroomen (2007) dans le sens où une suppression n'a été obtenue que lorsque les évènements étaient liés et prédictibles temporellement. Miki, Watanabe et Kagiri (2004) ont également réalisé une étude en MEG qui va dans ce sens. Les auteurs ont proposé deux types de stimuli aux participants : alors qu'un /a/ était prononcé, il était visuellement accompagné de 4 images (Filler > image 1 (800 ms) > image 2 (800 ms) > image 3 (400 ms)). Dans un cas, les images 1, 2 et 3 représentaient une locutrice bouche fermée. Dans l'autre cas, l'image 2 était celle de la locutrice prononçant un /a/, les images 1 et 3 étant les mêmes que précédemment décrites. La prononciation du /a/ durait 240 ms et était présentée en même temps que l'image 2. Les auteurs n'ont observé aucune modulation de la M100 auditive, cela confirmant à la fois

[34] Trad. « La capacité du cerveau à moduler sa sensibilité aux stimuli entrants ».

l'importance de la cohérence temporelle entre les informations auditives et visuelles, mais également que l'interaction audiovisuelle est causée par un filtrage sensoriel (qui ne peut avoir lieu ici, les deux informations étant présentées en même temps).

Ce filtrage aurait sans doute lieu de manière très précoce comme l'ont montré Lebib, Papo, de Bode et Baudonnière (2003). Les auteurs ont utilisé les voyelles /a, i, y, ø/ du français dans des présentations qui pouvaient être congruentes ou incongruentes, mais en faisant également varier la discriminabilité sur la base des informations visuelles, les voyelles /a/ et /i/ étant caractérisées par des propriétés visuelles très distinctives. Les auteurs ont observé un effet global d'amorçage des informations visuelles sur les informations auditives. Les stimuli incongruents impliquant /a/ et /i/ entraînent une P50 plus importante, ce qui serait le résultat d'une détection précoce de la non-congruence des entrées sensorielles, entrainant donc une augmentation de cette composante (*gate in*). Cela est cohérent avec l'absence de différence qu'ils observent pour les stimuli congruents facilement discriminables et les stimuli congruents difficilement discriminables. En effet, l'information est dans ces cas toujours redondante ce qui ne génère pas de modulation de l'onde entre ces deux conditions. Cependant, les conditions présentant des voyelles difficilement discriminables (les participants ne pouvant dans ces cas pas distinguer les évènements congruents des incongruents) ne modulent pas la P50, que les informations soient congruentes ou non. On peut supposer que, comme l'avaient observé Jesse & Massaro (2010) sur la N1, la réduction d'amplitude de l'onde est modulée par la quantité d'information visuelle disponible dans le signal. Il existe donc bien un filtrage intersensoriel précoce mis en évidence par la décroissance de la P50 pour les évènements sensoriellement redondants.

En résumé, les informations visuelles interagissent de manière précoce avec le cortex auditif (Besle et al., 2004 ; van Wassenhove et al., 2005). Une réduction de l'activité cérébrale est observée lorsque des mouvements anticipatoires sont disponible car ceux-ci permettent selon les auteurs soit d'anticiper les entrées acoustiques subséquentes (Besle et al., 2004 ; van Wassenhove et al., 2005), soit de filtrer les informations redondantes (Boutros, Barkerb, Tuetingb, Wub, & Nasrallahb, 1995 ; Boutros & Belger, 1999) ce qui permet d'alléger les traitements.

Alors que les processus neuronaux sous-jacents à l'intégration ne sont pas encore parfaitement compris, les données accumulées depuis une dizaine d'années ont permis de faire évoluer les modèles théoriques de l'intégration audiovisuelle.

2.1.3 MODELES D'INTEGRATION AUDIOVISUELLE

De nombreux modèles tentent d'expliquer comment et quand les informations auditives et visuelles sont combinées (voir (Boutros, Barkerb, Tuetingb, Wub, & Nasrallahb, 1995; Boutros & Belger, 1999) pour des revues). L'une des questions débattue dans la littérature concerne l'architecture de fusion audiovisuelle. Ainsi Klatt (1979), Massaro (1998) et McGurk & MacDonald (1976) ont postulé que les informations auditives et visuelles étaient traitées séparément, n'étant fusionnées qu'après la catégorisation phonologique. A l'opposé, les théories amodales de la perception audiovisuelle avancaient l'existence de mécanismes de fusion précoce qui prennent place avant même le niveau phonologique. Plus tardivement, Berthommier (2004) proposa même que la fusion soit précédée d'un niveau pré-phonétique, qui permet un « liage » entre les modalités. Cette dichotomie entre l'intégration précoce et tardive des signaux façonnera le cadre dans lequel nous nous placerons pour décrire les modèles de la perception audiovisuelle. Signalons toutefois que cette thèse n'a pas pour but de répondre aux questions spécifiques à l'intégration audiovisuelle dans les processus de perception de la parole; c'est la raison pour laquelle les modèles d'intégration ne seront présentés que de manière très succincte.

2.1.3.1 MODELES DE PERCEPTION DE LA PAROLE A INTEGRATION « TARDIVE »

Les modèles « *Lexical Access from Spectra and Face Parameters* », LASFP, (Klatt, 1979, cité par Klatt, 1989), « *Fuzzy Logical Model of Perception* », FLMP, (Massaro, 1998), « *Vision Place Audition Manner* », VPAM, (McGurk & MacDonald, 1976) supposent que l'intégration des informations visuelles et auditives se ferait après le traitement séparé des deux sources. Les partisans des théories auditives de la perception de la parole, même si celles-ci mettent l'accent sur la dominance de la modalité acoustique lors de la perception dans des conditions normales, considèrent les informations visuelles dans quelques cas. Diehl et Kluender (1989) défendaient l'idée que les deux informations sont prises en compte. Les informations visuelles sont catégorisées mais fusionnées seulement à un stade tardif du

traitement de l'information. Ils considèraient que ces associations sont le fruit d'un apprentissage implicite. Cependant et comme traité plus haut, de nombreux résultats, comme l'intégration chez des bébés (Rosenblum, Schmuckler, & Johnson, 1997), les singes (Sliwa, Duhamel, Pascalis, & Wirth, 2011), ou des intégrations impliquant des modalités différentes (i.e., toucher ; Fowler & Dekle, 1991) qui n'avaient pas été impactés par un apprentissage, vont à l'encontre de cette hypothèse d'associations apprises implicitement.

Le « *Fuzzy Logical Model of Perception* » (FLMP) de Massaro (1998, 2004) traite directement de la perception multimodale de la parole et postule que la perception audio-visuelle serait semblable aux mécanismes de reconnaissance de formes, même si cela demande de considérer des entrées sensorielles distinctes. Le traitement s'effectuerait en parallèle pour les différentes entrées sensorielles, évaluées séparément. Dans un premier temps, les entrées sensorielles, visuelles et auditives (ou issues d'autres modalités), seraient décomposées en propriétés ou « traits » (e.g., lors de la perception d'un /p/, un trait visuel sera par exemple le « degré d'ouverture labiale » et le « *burst* » sera un trait auditif) puis ces traits sont par la suite comparés à des prototypes phonétiques stockés en mémoire. Ce traitement serait donc effectué par des structures indépendantes les unes des autres. Une valeur est attribuée à chaque trait afin de quantifier l'adéquation entre le signal et le prototype. Cette valeur est comprise entre 0 et 1 (0.5 correspond à une ambiguïté complète, une valeur proche de 0 correspond à une inadéquation entre signal et prototype, et une valeur proche de 1 correspond à une bonne adéquation). Les codes extraits des informations auditives et visuelles seraient par la suite mis en commun afin de pouvoir les comparer au répertoire phonologique qui contiendrait des représentations auditives et visuelles des phonèmes connus. Le meilleur « candidat » sera ainsi sélectionné. Ce modèle suppose donc que les informations visuelles et auditives sont traitées séparément et ne sont mises en commun qu'après le traitement phonologique. De plus, ce modèle se base sur une loi fixe qui attribue un poids à des traits auditifs ou visuels, limitant donc l'impact de facteurs attentionnels, individuels ou contextuels.

2.1.3.2 MODELES DE PERCEPTION DE LA PAROLE A INTEGRATION « PRECOCE »

Les théories amodales de la perception audiovisuelle se situent dans la continuité de la Théorie Motrice de la Perception de la Parole (Liberman et al., 1967 ; Liberman & Mattingly, 1985 ; Liberman & Whalen, 2000) et de l'approche directe/écologique de la perception de la parole (Fowler, 1986). Les théories motrices ont l'avantage d'expliquer assez facilement

l'intégration audiovisuelle de la parole. Celles-ci postulent que les informations visuelles et auditives sont combinées précocement dans une représentation spatiale amodale (gestuelle). Cela vient du fait que ces théories sont dérivées de la proposition d'information indépendante de la modalité de Summerfield (1987). Les "*modality neutral theories*" supposent que la nature de ce codage ne serait pas dépendante de la modalité perceptive mais serait associée aux caractéristiques articulatoires du langage. Notre système coderait donc l'information en termes de gestes articulatoires, indépendamment de la modalité de perception (i.e., visuelle, auditive, tactile, etc). L'intégration étant considérée comme aveugle à la modalité d'entrée dès le début de la perception. Les études ayant montré une intégration également sur la base tactile (toucher la joue, les lèvres, ou le cou du locuteur), même lorsque les participants sont naïfs à ce genre de pratique, semblent indiquer une capacité d'intégration non pas basée sur les associations inter-modales apprises, mais bien sur une capacité sensorielle ancrée qui structure la parole et qui est amodalement dirigée (Sams, Möttönen, & Sihvonen, 2005 ; Treille, Cordeboeuf, Vilain, & Sato, 2014).

L'hypothèse d'une intégration dite "précoce" est également étayée par le fait que l'intégration des deux sources d'informations est automatique. Elle n'impliquerait aucun processus de haut niveau. L'expérience de Paris et al. (2013) vient appuyer cette idée. Ils présentent aux participants des séquences auditives (i.e., /ba/ et /da/) précédées de mouvements articulatoires congruents ou non au signal qui suit. Les résultats montrent qu'avec seulement 20% d'essais congruents (où le contenu visuel permettait de prédire le signal acoustique qui allait suivre), les participants obtiennent tout de même un avantage temporel grâce aux informations visuelles. Si l'intégration demandait des traitements non automatiques ou impliquerait des traitements de haut niveau, le nombre important de stimuli non-congruents devrait inciter les participants à ne pas se baser sur les mouvements labiaux (car cet indice n'est pas pertinent). Or, les participants se basent bien sur cette information alors que dans la majorité des cas, les mouvements labiaux ne prédisent pas le contenu auditif. Ces résultats rendent donc peu probable l'implication de processus cognitifs de haut niveau et atteste de l'effet d' « amorcage » des informations visuelles. Les données neurophysiologiques ont également fournies des arguments en faveurs de l'intégration précoce puisque des modulations des réponses neuronales sont observés dès 100 ms lors de présentations bimodales (Besle et al., 2004 ; van Wassenhove et al., 2005).

D'autres auteurs envisagent que le système de perception intègre à la fois des composantes perceptives et motrices (Remez, 1996, 2005 ; Schwartz et al., 2012). Le modèle de Schwartz et collaborateurs postule l'existence d'un mécanisme de « liage » qui interviendrait de manière précoce, avant que la fusion ne soit réalisée. Le liage permettrait de regrouper les différents types d'information présents dans la parole, qu'elles soient auditives, visuelles, tactiles, etc. Ce mécanisme permettrait de rendre compte des résultats suggérant que la fusion entre les informations auditives et visuelles n'est pas automatique ; elle serait en outre modulée par des facteurs attentionnels (Alsius, Navarra, Campbell, & Soto-Faraco, 2005), contextuels (Nahorna, Berthommier, & Schwartz, 2010) et interindividuels (Schwartz, 2010).

L'hypothèse d'une intégration précoce a également été confortée par des études en neurophysiologie. Les régions cérébrales sensibles à la parole auditive (e.g., cortex auditif primaire) répondent également à la parole visible (Musacchia, Sams, Nicol, & Kraus, 2005) et ce, de manière précoce (Wassenhove et al., 2005). L'intérêt pour la localisation des sites impliqués lors de l'intégration audiovisuelle a permis de mieux comprendre le décours temporel de cette intégration.

2.1.3.3 SITES NEURONAUX DE L'INTEGRATION

Les sites neuronaux d'intégration audiovisuelle, dans le cadre de la parole, ont été largement examinés à partir des années 2000. Dans une étude de Calvert, Campbell et Brammer (2000) en IRMf, les participants devaient suivre et comprendre une histoire lue à voix haute par un locuteur dont seule la partie inférieure du visage était visible sur l'écran. L'histoire pouvait être présentée auditivement, visuellement ou dans la double modalité. Par ailleurs, les participants étaient soumis à deux sessions expérimentales : dans la première, le mouvement des lèvres correspondait au texte entendu (messages congruents), tandis que dans la deuxième session, les deux signaux correspondaient à deux histoires différentes (messages incongruents). Comme dans leur étude sur la perception passive de stimuli audiovisuels « non langagiers » (Calvert, 2001), seuls les sites pour lesquels une augmentation de la réponse pour des stimuli congruents et de diminution de la réponse pour des stimuli incongruents étaient considérés comme des sites d'intégration. Seul le sillon temporal supérieur (STS) gauche répondait à cette double contrainte. D'autres régions ne répondaient qu'au critère d'augmentation pour les stimuli congruents : les cortex sensoriels spécifiques, le

gyrus frontal médian gauche et le lobule pariétal droit. Les auteurs ont suggéré que les effets dans les cortex sensoriels spécifiques étaient dus à des activations de feedback à partir des sites intégratifs tels que le STS.

Ces auteurs ont également observé que la quantité d'activation en condition audiovisuelle était supérieure à la somme des activités mesurées en conditions unimodales auditive seule (A) et visuelle seule (V). Ce pattern supra-additif (AV > A + V) a également été obtenu en comportement (Navarra & Soto-Faraco, 2007 ; Schwartz et al., 2004) (cf. section 1.3.1.2). Il a été interprété comme le résultat de l'intégration entre les informations auditives et visuelles. *A contrario*, lorsque les informations auditives et visuelles étaient incongruentes, un pattern sub-additif a été retrouvé (i.e., AV < A + V). Pour ces raisons, Calvert et al. (2000) ont identifié le STS comme étant une structure cérébrale jouant un rôle clé dans le processus de fusion de ces deux informations.

En accord avec cette idée, des expériences de stimulation du STS (via la technique TMS ou stimulation magnétique transcrânienne), montrant les interférences créées lors de l'intégration multimodale, semblent indiquer son rôle lors de la fusion (Beauchamp, Nath, & Pasalar, 2010). Les auteurs observent en effet une interférence lors de la présentation de stimuli incongruents de type McGurk, mais pas avec le traitement de stimuli audiovisuels de parole congruents. Ces résultats viennent donc conforter l'idée que le STS jouerait un rôle critique dans l'intégration audiovisuelle des signaux de parole. De nombreux résultats semblent converger dans ce sens (Callan et al., 2003 ; Miller & D'Esposito, 2005 ; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003). Beauchamp, Lee, Argall et Martin (2004) précisent en outre que des portions différentes (1 à 2 mm) du STS reçoivent des stimulations auditives et visuelles qui sont ensuite transmises vers une troisième partie « multisensorielle ».

Cependant, le rôle du STS dans l'intégration audiovisuelle sera par la suite remis en question (voir Hocking & Price, 2008 pour une critique). Cette zone serait bien impliquée lors de la perception multimodale, mais alors qu'il était communément admis que l'intégration était un processus de haut niveau engageant les cortex hétéro-modaux associatifs, des travaux récents ont bouleversé cette vision et suggèrent que l'interaction entre les modalités pourrait être en réalité gérée par les cortex unimodaux de bas niveau. Chez l'animal, Schroeder & Foxe (2005) ont montré des interactions multi-sensorielles dans le cortex auditif primaire. Chez l'humain, des études en MEG ont permis de montrer que l'interaction qui a lieu dans le

cortex auditif lors de la présentation de stimuli multimodaux précède l'activation du STS (qui se situe autour de 250-600 ms ; Möttönen, Schürmann, & Sams, 2004). Dans ce sens, van Wassenhove et al. (2005) et Besle et al. (2004) obtiennent des données en EEG indiquant une influence des informations visuelles à des étapes précoces du traitement (100-200 ms).

Olson, Gatenby et Gore (2002) penchent pour un rôle plus large des gyrus et sillon temporaux supérieurs qui seraient davantage impliqués dans l'analyse des mouvements biologiques que dans l'intégration bimodale. Ces chercheurs ont mené une étude sur dix volontaires en IRMf, avec deux conditions uniquement visuelles et deux conditions audiovisuelles non congruentes (type McGurk), dont l'une présentait un signal sonore désynchronisé. Les stimuli étaient prononcés par une locutrice, et ce dans quatre conditions : visuelle statique, visuelle animée (présentant une vidéo avec un mot toutes les trois secondes), audiovisuelle synchronisée et audiovisuelle désynchronisée. Les régions spécifiquement activées par les propriétés visuelles dynamiques (comparaison des deux conditions visuelles (animée vs. statique)) sont l'homologue droit de l'aire de Broca, le gyrus temporal supérieur gauche et les gyrus et sillon temporaux supérieurs droits. Les auteurs précisent qu'en élevant le seuil statistique ($p < 0.01$) seuls les gyri temporaux supérieurs droit et gauche restent activés. Le résultat le plus intéressant en ce qui nous concerne est issu de la comparaison audiovisuelle qui montre une activation de la région du claustrum gauche (fine couche de matière grise située aux abords de l'insula). Ces résultats, selon les auteurs, plaident en faveur de l'idée selon laquelle les sites de traitements unimodaux seraient aussi les sites d'intégrations bimodaux via des relais sous corticaux. Une étude en EEG semble confirmer la médiation de l'intégration par des structures sous corticales (Musacchia, Sams, Nicol, & Krauss, 2006).

D'autres chercheurs avancent que ce serait les zones motrices liées à la planification et à l'exécution qui seraient responsables de cette intégration précoce. Des zones comme l'aire de Broca, le cortex pré-moteur et l'insula antérieure seraient impliquées à cause des neurones miroir. En effet, l'aire de Broca est connue pour être l'homologue chez le macaque du cortex pré-moteur inférieur qui est le siège des neurones miroir qui répondent à l'action ou perception dirigée vers un but de la main ou de la bouche (Rizzolatti & Craighero, 2004). Ils serviraient donc lors de l'imitation et aideraient à l'action et à la compréhension. Dans ce sens, l'aire de Broca est activée durant de la lecture labiale silencieuse (R. Campbell et al., 2001), et l'écoute passive (Wilson et al., 2004). D'autres travaux vont dans le sens de

l'implication de l'aire de Broca également lors du décodage d'un signal de parole en modalité audiovisuelle (Wilson et al., 2004). En effet, Callan et al. (2003) ont par exemple montré que le cortex pré-moteur et l'aire de Broca étaient plus fortement activés lors de la perception de stimuli audiovisuels que lors de la perception de stimuli auditifs seuls. Ces régions étant impliquées dans la planification et l'exécution de la production de la parole, cela suggère que la perception de l'information visuelle en présence de l'information auditive pourrait être supportée (du moins en partie), par les régions motrices responsables de la production du langage oral.

Même si les sites neuronaux de l'intégration audiovisuelle ne sont pas encore clairement identifiés, nous savons aujourd'hui que l'intégration met en jeu des traitements de bas niveau et que celle-ci a lieu de manière très précoce.

2.2 PERCEPTION AUDIOVISUELLE DES LANGUES ETRANGERES

La plupart d'entre nous a déjà pu réaliser que tenir une conversation dans une langue étrangère est difficile. Cela est encore plus difficile quand notre interlocuteur n'est pas physiquement présent, par exemple lorsque nous tenons une conversation téléphonique. Cela peut s'expliquer par le fait qu'écouter quelqu'un parler dans une langue étrangère revient à de l'écoute dans des conditions adverses (Arnold & Hill, 2001). Lorsque nos connaissances dans une langue sont limitées, nous ne pouvons pas exploiter efficacement le signal auditif, celui-ci étant perçu au travers du crible de la langue maternelle. Dans le cadre de la langue maternelle, lorsque les conditions d'écoute sont bruitées nous pouvons nous appuyer sur les mouvements oro-faciaux du locuteur afin améliorer notre compréhension (Benoît et al., 1994 ; Binnie, Montgomery, & Jackson, 1974 ; Calbour & Dumont, 2002 ; Erber, 1969; Lallouache, 1991 ; Middelweerd & Plomp, 1987 ; Mohamadi & Benoit, 1992 ; cf. 2.1.1.3.1 Présentation audiovisuelle dans le bruit). Est-il possible d'en faire autant lors de la perception d'une langue étrangère et plus particulièrement de phonèmes qui ne sont pas répertoriés dans notre inventaire phonologique ? La question que nous posons est de savoir si les difficultés que nous rencontrons pour interpréter des phonèmes qui nous sont inconnus peuvent elles aussi être modulées par l'utilisation des mouvements articulatoires visibles afin de désambiguïser le signal acoustique.

2.2.1 DEVELOPPEMENT

2.2.1.1 QUAND LA SURDITE PHONOLOGIQUE REND « AVEUGLE »

Comme nous l'avons vu précédemment, les bébés utilisent très précocement l'information visuelle. Ils sont par exemple capables sur la seule base de ces informations de discriminer deux histoires énoncées en français ou en anglais à l'âge de six mois (Weikum et al., 2007). Les enfants ont également très tôt la capacité de faire correspondre de la parole non-native avec la réalisation articulatoire qui convient. Ceci a été mis en évidence aussi bien par l'analyse de temps de fixation (Pons et al., 2009) que des taux de succion (Walton & Bower, 1993). Pons et al. (2009) ont présenté à des enfants hispanophones et anglophones de 6 et 11 mois le contraste phonologique /b-v/ (qui n'existe pas en espagnol mais qui existe en anglais) dans des séquences CV ou V était toujours un /a/. La présentation des informations auditives et visuelles était séquentielle (Figure 11).

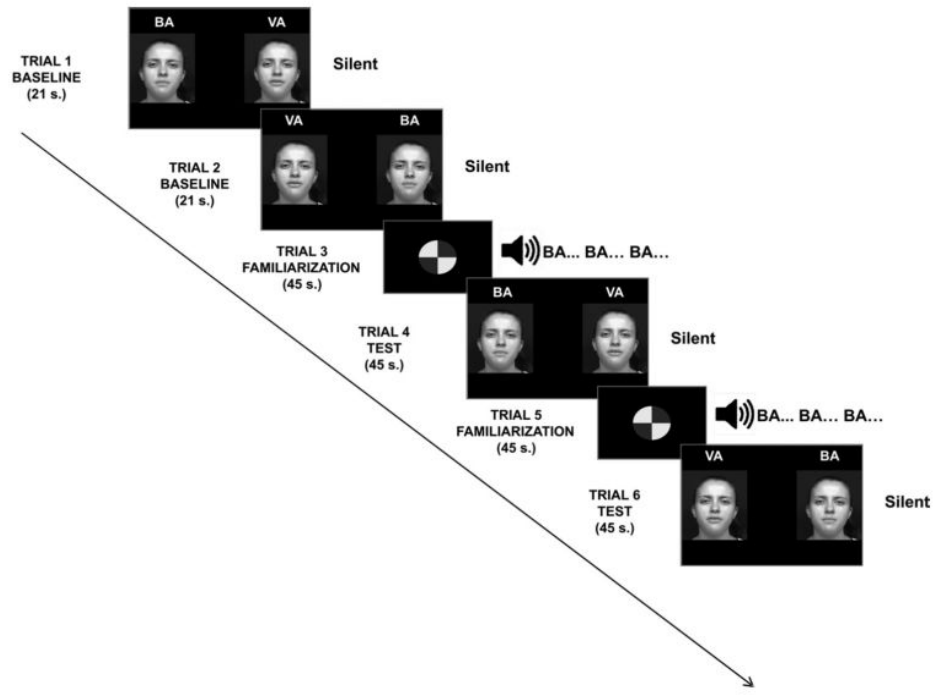

Figure 9. Représentation schématique de la procédure séquentielle utilisée par Pons et al. (2009).

Dans un premier temps, les enfants étaient d'abord familiarisés aux deux articulations silencieuses. Puis une des deux syllabes était répétée plusieurs fois (i.e., phase de familiarisation). Une seconde plus tard, deux visages articulant chacun une syllabe étaient présentés côte-à-côte (i.e., phase test). Si les enfants ont à la capacité de faire correspondre la syllabe avec son articulation visible, ils devraient alors regarder préférentiellement la vidéo

correspondant à la syllabe précédemment entendue. A six mois, en cohérence avec les résultats obtenus lors de présentation unimodale, les deux phonèmes sont discriminés par les deux populations. Les enfants sont donc capables de faire correspondre la syllabe /va/ entendue lors de la familiarisation avec l'articulation silencieuse présentée quelque secondes plus tard et de la discriminer de l'articulation concurrente /ba/. Cependant, comme c'est le cas pour la perception auditive seule, à 11 mois, les bébés hispanophones ne parviennent plus à distinguer /b/ de /v/ alors que les anglophones y restent sensibles (Figure 12).

Dans la même veine, les enfants sont sensibles à l'asynchronie entre son et image indépendamment de la familiarité avec la langue dans laquelle un discours est énoncé (Pons &

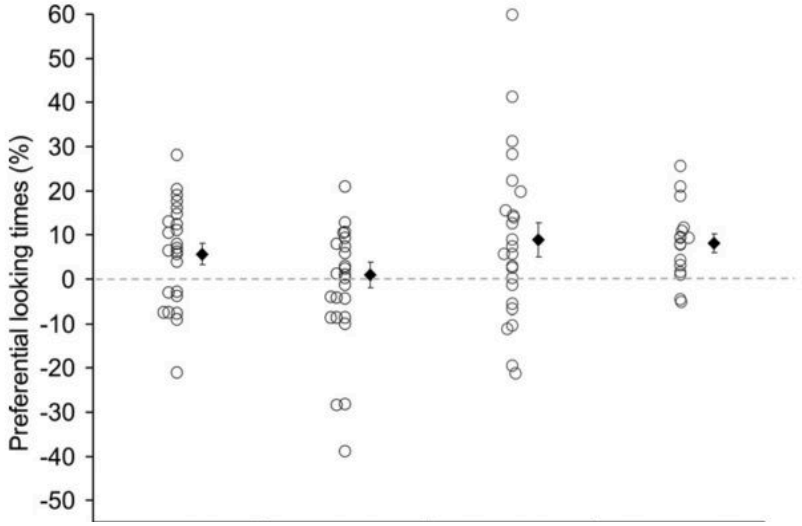

Figure 10. Distribution des différences entre le pourcentage du temps totale pendant lequel l'enfant a regardé le visage correspondant à la syllabe présentée visuellement moins le pourcentage du temps total durant lequel ils ont regardé le visage correspondant à la syllabe présentée visuellement durant les essais de baseline, pour chaque groupe d'âge (6 et 11 mois) et langage (anglais et espagnol). Les cercles pleins représentent la différence moyenne. Extrait de Pons et al. (2009).

Lewkowicz, 2014), ce qui n'est plus le cas après 10-12 mois. L'affinage perceptif fait donc son œuvre même lors d'une présentation bimodale. Les résultats précédents attestent en effet de l'incapacité des bébés à exploiter des informations visuelles des phonèmes ou séquences exprimés dans une langue qui n'est pas familière. Cependant, les résultats ne peuvent s'expliquer que par une sous-exploitation des informations visuelles. En effet, les enfants restent capables de distinguer visuellement des contrastes qui existent dans leur langue maternelle. Pourquoi sont-ils donc incapables de décoder la parole visuelle dans une langue qui n'est pas parlée dans leur environnement ?

2.2.1.2 L'HYPOTHESE DU MANQUE DE SAILLANCE DES INDICES VISUELS NON-NATIFS

Lewkowicz & Hansen-Tift (2012) ont observé que les enfants de huit mois regardent la bouche du locuteur, alors qu'ils tournent préférentiellement leur regard vers les yeux à 12 mois. Cependant, lorsqu'un locuteur s'exprime dans une langue étrangère (e.g., l'espagnol), les enfants tournent à nouveau leur regard vers la bouche. Les auteurs l'expliquent par un besoin de trouver une source d'information supplémentaire et/ou redondante afin de compléter le signal auditif, comme le font les adultes (Vatikiotis-Bateson et al., 1998). Cependant, d'autres explications peuvent être envisageables, comme une simple réaction à la nouveauté. Quelle qu'en soit l'explication, les bébés exposés à de la parole non-native ont un comportement différent de celui observé lorsqu'ils perçoivent la langue maternelle, et orientent leur regard vers la bouche. Cependant et comme l'attestent les études ci-dessus, il semblerait qu'ils ne soient pas, pour le moment du moins, capables d'utiliser l'information fournie par le canal visuel lorsque le signal n'est pas natif. L'hypothèse du manque de saillance des indices pertinents, notamment des informations non-linguistiques, a alors été avancé pour expliquer ces difficultés.

Dans ce sens, Ostroff (2000) a voulu tester l'impact de l' « *infant-directed-speech* » (ID) sur les capacités de discrimination des enfants. Ce type de discours est, comme son nom l'indique, une façon particulière qu'ont les adultes de s'exprimer lorsqu'ils parlent à des enfants, surtout lorsqu'ils sont bébés. Ce discours vise à accentuer certaines caractéristiques par une *hyper-articulation*, accompagnée notamment d'une augmentation de la hauteur. Les enfants montrent plus d'attention pour ce type de discours que pour le « *adult-directed speech* » (AD) (Cooper & Aslin, 1990). Ostroff (2000) inclue donc dans cette étude des syllabes de l'hindi (le contraste des occlusives rétroflexe et dentale) enregistrées soit en ID, soit en AD. Les deux types de séquences ont été présentés à des enfants de 11 mois dans deux conditions. La première montre le visage du locuteur et dans la seconde, des formes géométriques ont remplacé le visage, ceci dans le but de savoir si ce sont les indices visuels qui permettent la discrimination ou simplement l'accentuation des indices auditifs fournis par l'ID). L'auteur observe que les enfants ne sont pas capable de discriminer des phonèmes non-natifs prononcés en AD que le visage du locuteur soit présent ou non (Figure 13). En revanche ils augmentent significativement leur temps de fixation par rapport à la phase d'habituation quand les séquences étaient prononcées en ID et que les mouvements articulatoires étaient disponibles.

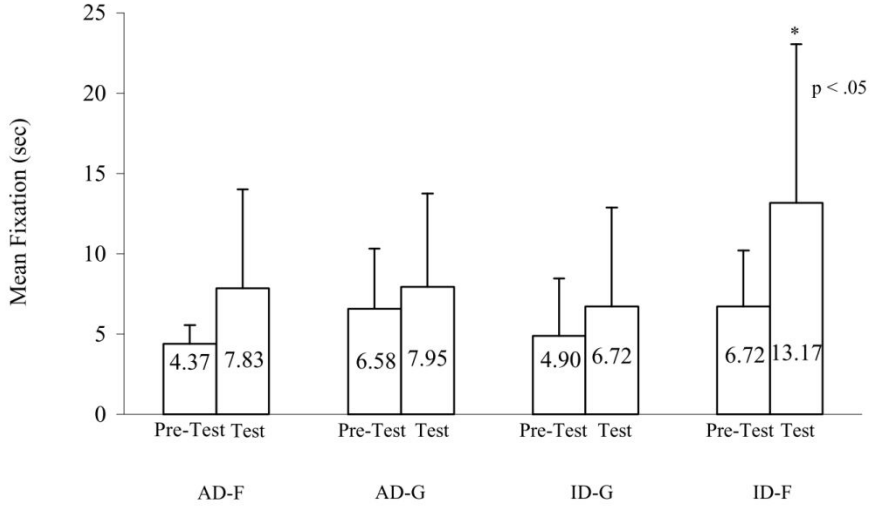

Figure 11. Temps de fixation moyen des enfants durant les phases de pré-test et de test en fonction de la condition de présentation: F = visage; G = Forme géométrique. (Extrait de Ostroff, 2000)

L'augmentation des temps de fixation est donc liée à l'exploitation des gestes articulatoires, qui lorsqu'ils sont plus marqués, peuvent être utilisés par les enfants afin de discriminer les syllabes non-natives. Il est donc envisageable que l'échec des enfants à discriminer des phonèmes qu'ils ne connaissent pas soit en partie dû à une saillance moindre des indices visuels avec lesquels ils ne sont pas familiers. Les mouvements oro-faciaux sont peut-être plus à même d'être exploité à l'âge adulte, sans pour autant que ceux-ci nécessitent d'être saillants.

2.2.2 UTILISATION DES INFORMATIONS ARTICULATOIRES LORS DE L'APPRENTISSAGE D'UNE LANGUE ETRANGERE CHEZ L'ADULTE

A l'âge adulte, nous savons utiliser les informations visuelles fournies par les mouvements oro-faciaux dans le but de distinguer deux langues, n'en connaître qu'une des deux suffit pour les différencier (Soto-Faraco et al., 2007). Cela n'est pas le cas si les deux langues sont inconnues (e.g., lorsqu'un anglais doit différencier un mot espagnol d'un mot catalan). Cependant, ces indices fournis par la vision, associés à des indices auditifs, sont, comme nous l'avons vu très riches en informations. Toutefois, ces mouvements sont-ils utilisables dans le cadre de l'apprentissage de langues étrangères, notamment lorsqu'aucune connaissances préalable n'a été acquise ? La littérature s'est arrêtée de manière privilégiée sur le cas des « apprenants », qui sont déjà familiarisés avec la langue grâce à l'acquisition de connaissances

sur la phonologie, le vocabulaire, les règles syntaxiques ou grammaticales de la langue. Le cas des individus monolingues qui débutent leur apprentissage de la langue étrangère est bien moins renseigné. Il s'agit d'examiner si l'ajout de la modalité visuelle dès le début de l'apprentissage pourrait limiter le phénomène de surdité phonologique, et donc faciliter l'apprentissage. Comme nous l'avons déjà signalé, le décodage phonologique est la première étape qui permet de comprendre notre interlocuteur, et d'elle dépend la suite des traitements de la parole. Nous allons dans une premier temps passer en revue les études menées sur des participants qui débutent leur apprentissage d'une langue étrangère pour continuer avec les travaux réalisés chez les apprenants au niveau intermédiaire ou expert.

2.2.2.1 UTILISATION DES INFORMATIONS VISUELLES PAR LES DEBUTANTS

La plupart des difficultés que rencontrent les débutants dans une langue sont liés à l'influence des catégories phonologiques de la langue maternelle. La surdité phonologique « *ensures that non-native speakers will perceive at least some L2 vowels and consonants differently than do native speakers*[35]» (Flege, 1995, pp. 237). Les phonèmes de la L2 sont cartographiés selon les représentations de notre langue qui sont les plus proches acoustiquement ou articulatoirement (Soto-Faraco et al., 2007). L'apprentissage de contrastes complexes, comme par exemple /s-θ/ pour les francophones natifs, est difficile et demande du temps car il inclut un phonème *nouveau*. Ce phonème sera donc au début de l'apprentissage perçu comme un allophone d'un des phonèmes de la langue maternelle, dans ce cas /s/ (Flege, 1995). Le phonème /θ/ sera assimilé à /s/, ce qui va provoquer des confusions entre des mots comme *sick* et *thick* (Berger, 1951 ; Brannen, 2002).

Selon le *Speech Learning Model* de Flege (1995), lors de l'acquisition de la langue maternelle, les enfants doivent trouver des indices perceptifs suffisants pour identifier les sons et pouvoir les classifier. Durant cet apprentissage, les informations visuelles jouent un rôle très important dans la mise en place des catégories phonologiques natives (Mills, 1987). Ces mécanismes seraient toujours actifs à l'âge adulte et seraient employés lors de l'apprentissage de la L2. Les individus confrontés à une nouvelle langue pourraient donc, comme l'enfant, bénéficier de l'information visuelle, par exemple pour désambiguïser la place d'articulation,

[35] Trad. « garantie que les locuteurs non natifs percevront au moins quelques voyelles et consonnes de la L2 de manière différente que ne les percevra une locuteur natif ».

comme pour /m/ et /n/. En effet, lorsque nous sommes face à une langue étrangère nous avons tendance à porter plus d'attention sur la région buccale (Hayashi & Sekiyama, 1998 ; Werker, Frost, & McGurk, 1992). L'auditeur pourrait ainsi détecter visuellement des détails phonétiques qui diffèrent entre ces sons. Cela permettrait à l'apprenant d'une langue étrangère de mieux comprendre les différences articulatoires entre les phonèmes de sa langue maternelle et ceux de celle qu'il apprend. Ainsi deux sons qui sont assimilés lors de la présentation auditive pourraient bénéficier des différences articulatoires visibles pour être discriminés et placés dans une nouvelle catégorie. C'est la conclusion de Wang et al. (2008) qui indiquent que les apprenants devraient se focaliser « *on L2 specific visual speech cues to establish correct L2 categories*[36] » (pp. 1725). L'utilisation des informations visuelles même sans connaissance préalable de la langue permettrait à la fois une meilleure compréhension de ses phonèmes mais aussi permettre de limiter dès le début de l'apprentissage les mauvaises catégorisations, facilitant la suite de l'apprentissage.

Les études concernant l'apport des informations visuelles pour les personnes n'ayant aucune notion préalable dans une langue étrangère sont peu nombreuses. En 1999, Davis et Kim ont mis en évidence que des anglophones ne connaissant pas le coréen parvenaient mieux à détecter un phonème coréen inséré dans une phrase porteuse lorsqu'ils avaient la possibilité d'utiliser les informations visuelles même sans connaissance préalable d'une langue. Ils ont également testé la capacité d'australiens anglophones à répéter des phrases courtes en coréen (Davis & Kim, 1998, 2001). Dans la première phase, les participants étaient exposés à une vidéo de la locutrice. Pour la moitié des séquences vidéo, seule la partie supérieure du visage (i.e., au-dessus du nez) était présentée. Pour l'autre moitié des séquences, c'était la partie inférieure du visage qui était présentée (i.e., en dessous du nez). La tâche était de répéter les phrases prononcées. Dans une seconde phase, des phrases étaient prononcées et les participants devaient dire s'ils avaient déjà entendu ces phrases dans la première phase. Les résultats suggèrent que la précision des répétitions des phrases ainsi que le nombre de syllabes rappelées dans la première phase était meilleure lorsque les sujets avaient vu la partie inférieure du visage que lorsqu'ils avaient vu la partie supérieure de ce visage. Cette expérience révèle également que les phrases qui bénéficient le plus de l'information visuelle sont celles qui contenaient le plus de mouvements de lèvres. Concernant le nombre moyen de

[36] Trad. « se focaliser sur les indices visuels spécifiques à la L2 pour établir des catégories de L2 valides ».

phrases reconnues dans la seconde phase, celui-ci a été plus grand lorsque les participants avaient perçu la partie inférieure du visage que lorsqu'ils avaient vu la partie supérieure.

Une autre étude propose des résultats en la défaveur de l'utilisation des informations visuelles pour désambiguïser un contraste non-natif. Dans l'étude de Pons et al. (2009) que nous avons décrite dans le Chapitre 3.2.1.1. « Quand la surdité phonologique rend aveugle », il y avait deux groupes d'adultes hispanophones et anglophones qui devaient discriminer en présentation audiovisuelle du contraste /b-v/ (qui existe en anglais mais pas en espagnol). Les résultat indiquent que les hispanophones n'étaient pas capables d'exploiter les informations visuelles fournies par les mouvements labiaux lorsque le phonème ne faisait pas partie de leur répertoire phonologique (i.e., /v/), malgré leur familiarité avec le visème de la labiodentale /f/ qui est le même que /v/ et qui existe en espagnol. Cependant, il est bon de noter que le mode de présentation des différentes modalités était séquentiel dans cette étude. En effet, dans leur paradigme, un stimulus auditif était répété deux fois, puis après une seconde, les participants devaient choisir laquelle des deux vidéos présentées correspondait à l'articulation de la syllabe qu'ils avaient entendu précédemment. Or si les hispanophones ne sont pas capables de percevoir, en modalité auditive seule, la différence entre /va/ et /ba/, il est logique que la catégorisation subséquente soit impossible.

Les études menées chez les apprenants ont révélés des résultats assez partagés. Alors que certaine études attestent d'un avantage lors de la présentation audiovisuelle de phonème non-natifs, d'autre peines à observer un avantage.

2.2.2.2 UTILISATION DES INFORMATIONS VISUELLES PAR LES APPRENANTS

Pour des apprenants, des études ont permis de mettre à jour un avantage lié à la présentation audiovisuelle en *production*. L'étude de Reisberg, McLean et Goldfield (1987) atteste de l'impact de la présentation audiovisuelle dans une tâche de répétition. Ils mettent en évidence que des anglophones apprenant le français répètent mieux des phrases en français lorsqu'on leur présente la vidéo du locuteur en train de les prononcer, que lorsque seul le son est présenté. Cette étude comprenait également un versant perception. Là encore, la présentation audiovisuelle semble bénéfique pour comprendre des phrases au contenu sémantique complexe (« La critique de la raison pure » de Emmanuel Kant !). Ces résultats ont

notamment été répliqués à l'aide d'un autre paradigme par Arnold et Hill (2001) qui obtiennent des avantages à la fois pour des phrases au contenu sémantique et syntaxique complexe mais également dans le cadre de phrases prononcées avec un accent inconnu. Il semblerait donc que les informations visuelles entraînent une amélioration des performances en répétition mais également un avantage perceptif. Des études se sont focalisées sur la perception de syllabe ou de phonème non-natifs.

En 2002, Ortega-Llebaria, Faulkner et Hazan ont testé des hispanophones et des anglophones natifs sur leur capacités à identifier 9 consonnes de l'anglais (i.e., /b d g p t k v z ð f s ʃ t ʃ d ʒ m n/) et de 7 monophtongues et 2 diphtongues. Ils observèrent une amélioration des performances de 3.7% lors de la présentation audiovisuelle de consonnes. La présentation audiovisuelle des voyelles améliorait les performances des hispanophones de 1.7%, ce qui n'était pas significatif. L'ajout des informations visuelles permettait cependant aux deux populations de réduire les erreurs de place d'articulation pour les trois modes d'articulation. Malgré cet avantage lors d'une présentation audiovisuelle, une analyse plus poussée à amené les auteurs à conclure que la réduction des confusions induites par les informations visuelles concernait les consonnes qui faisaient partie du répertoire phonologique des hispanophones, les confusions langage-dépendante n'étant pas réduites par l'ajout des informations visuelles (e.g., /b/-/v/). Il semblerait donc à première vue que les relations entre les systèmes phonologiques de la L1 et de la L2 affectent l'utilisation des indices visuels de la L2.

Suite à ces premiers résultats sur un large panel de consonnes, qui entraient en contradiction avec ceux de Hardison (1999) qui elle obtenait une amélioration de 10% des performances de reconnaissance du contraste /r/-/l/ anglais lorsqu'il était perçu en modalité audiovisuelle par des japonais et des coréens, Hazan décidât avec Anke Sennema et Andrew Faulkner (2002) de réduire les contrastes étudiés à deux, afin de manipuler l'information fournie par les indices visuels. Les participants hispanophones devaient identifier trois consonnes de l'anglais (i.e., /p/-/b/-/v/). Le contraste /b/-/v/ était censé poser problème aux hispanophones qui ne possèdent pas la labiodentale fricative voisée, mais également le contraste /b/-/p/ puisque la réalisation du /p/ espagnol comporte des caractéristiques du /b/ anglais. Ils sont en effet tous deux réalisés comme des plosives aspirées et ont des VOT similaires. Les participants étaient exposés à des séquences de type CV, VCV ou VC présentées en modalité auditive, audiovisuelle et visuelle seule. Les résultats, même s'ils montrent un effet de la modalité de présentation, n'attestent pas d'une amélioration des performances lors de la présentation audiovisuelle. Les résultats dans la condition auditive

étaient similaires à ceux obtenus lors de la présentation audiovisuelle. L'utilisation des informations visuelles semblent donc compromise à partir du moment où le locuteur est étranger. Notons que les auteurs précisent tout de même qu'un bon nombre de participants, ceux qui avaient un niveau assez élevé de maîtrise de l'anglais, parvenaient à exploiter les informations visuelles pour désambigüiser le signal. A l'opposé, Hardison (2003) met à jour un avantage lié à l'utilisation des information visuelles par des japonais percevant /r/-/l/, après un entraînement audiovisuel (voir Hazan, Sennema, Iba et Faulkner (2005) pour des résultats contradictoires). Elle montre également un avantage de la présentation bimodale lors du pré-test (i.e., sans entraînement) alors que Hazan n'en trouvait pas pour ce contraste en 2002.

En 2005, Hardison mène une utilisant un protocole de *gating* qui amène des éléments intéressants, notamment sur le fait que l'identification de mot soit plus rapide lors d'une présentation audiovisuelle, qu'il s'agisse de mots imbriqués dans une phrase scriptée[37] ou spontanées. Il semblerait que la présentation audiovisuelle amène également un avantage temporel, comme c'est le cas pour les phonèmes natifs. Elle montre également que les mots qui commencent avec les phonèmes les plus difficiles à identifier étaient ceux qui bénéficiaient le plus de l'information visuelle, à savoir /θ/-/w/-/ɹ/ et /r/-/l/ en fonction du locuteur. Enfin, Wang et al. (2008) ont réalisé une étude dans laquelle des participants parlant le coréen, le mandarin et l'anglais percevaient des fricatives qui existaient ou non dans leur langues respectives. Les résultats montrent qu'ils utilisent l'information visuelle ce qui améliorait leur performance. Ils répliquèrent ces résultats en 2009 sur une population similaire. Les scores des participants parlant le mandarin passaient de 50 à 67% d'identification correcte lors de la perception de audiovisuelle de l'interdentale non-native. Pour ceux parlant le coréen, les scores passaient de 67 à 84% d'identification correcte pour la perception du même phonème non-natif. Ces résultats indiquent également une modulation liée spécifiquement aux langues des populations. Alors qu'aucune des deux populations ne possèdent la fricative interdentale dans leur répertoire phonologique, il semblerait que les coréens se basent plus sur les informations auditives alors que ceux parlant le mandarin s'appuient sur les informations visuelles. Cela est dû à la structure de leur répertoire phonologique.

[37] La différence entre la parole claire/scriptée/lue (la terminologie diffère en fonction des études (Prieto, 2004) et conversationnelle/ non scriptée se joue avant l'enregistrement des séquences. Pour le discours clair, le locuteur reçoit la consigne de s'exprimer comme s'il parlait alors que les conditions d'écoute sont bruitées ou que la personne à qui il s'adresse est légèrement sourde, ainsi il articule de manière plus soigneuse. Dans le cas de la parole conversationnelle, il doit s'exprimer de manière normale.

Globalement, on observe des divergences dans les résultats obtenus dans la littérature quant à l'utilisation de l'information visuelle. Alors que certaines études ont montré un impact de ces indices pour l'identification des contrastes non-natifs chez les apprenants (Davis & Kim, 2004 ; Erdener & Burnham, 2005 ; Hardison, 1999, 2003, 2005 ; Hayashi & Sekiyama, 1998 ; Hazan et al., 2006 ; Kluge, Reis, Nobre-oliveira, & Bettoni-techio, 2006 ; Sekiyama, Kanno, Miura, & Sugita, 2003 ; Thompson & Hazan, 2007 ; Wang et al., 2008, 2009), d'autres n'ont pas mis à jour d'avantage (Hazan, Sennema, & Faulkner, 2002 ; Hazan, Sennema, Iba, & Faulkner, 2005 ; Ortega-Llebaria, Faulkner, & Hazan, 2001). Ces différences dans les résultats nous poussent à nous interroger sur les facteurs qui impactent l'utilisation des informations visuelles présentent dans les mouvements oro-faciaux lors de la production de phonèmes non-natifs. Hazan et al. (2006), nous éclairent sur certains de ces facteurs. Ils concluent que « *visual similarity between the L1 and L2 is therefore a factor that may affect L2 phoneme acquisition. Two further factors, which are more language general are the relative weighting of acoustic and visual cues in L1: Learners who attach greater weight to visual information in L1 are more likely to attend to visual cues to phoneme contrasts in the L2. Finally, just as acoustic salience may be a factor in the ease of acquisition of the contrast, the visual salience of the contrast is likely to affect the degree to which learners attend to visual cues*[38] » (p.1750). Nous allons dès à présent détailler ces facteurs.

2.2.3 FACTEURS QUI MODULENT L'UTILISATION DES INFORMATIONS VISUELLES DE LA LANGUE ETRANGERE

La saillance perceptive des visèmes étudiés, le nombre de contraste visuels distinct dans la langue maternelle, les relations entre les phonèmes de la langue maternelle et la L2 en termes de contraste visuel, le poids relatif donné aux deux types d'informations en fonction de la langue perçue mais également de la langue du participant influencent la façon dont les mouvements articulatoires peuvent être exploités.

[38] Trad. « la similarité visuelle entre la L1 et la L2 est par conséquent un facteur qui peut affecter l'acquisition de phonème de la L2. Deux facteurs supplémentaires, qui sont langagiers, sont le poids relatif des indices acoustiques et visuels dans la L1 : les apprenants qui donnent le plus de poids aux information visuelles dans la L1 sont plus susceptible de faire attention aux indices visuels lors de la perception de phonème contrastifs de la L2. Enfin, tout comme la saillance acoustique peut être un facteurs dans la facilité d'acquisition d'un contraste, la saillance visuelle du contraste est susceptible d'affecter le degré avec lequel l'apprenant fait attention aux indices visuels ».

2.2.3.1 LA SAILLANCE

Les visèmes de la L2 sont impactés par la saillance visuelle au même titre que ceux de la langue maternelle. Cet impact a également été montré par (Dodd, 1977; Hardison, 1999; Hazan et al., 2006, 2005; Wang et al., 2008). Hazan et al. (2006), a observé que l'effet de la présentation des mouvements articulatoires variait en fonction de leur saillance perceptive. Ils testaient les capacités de japonais et d'espagnol à identifier le contraste entre les labiales /b/-/p/ et la labiodentale /v/ qui constitue des phonèmes très saillants visuellement ainsi que le contraste /r-l/ qui est moins visible. Leurs résultats revèlent un avantage audiovisuel faible mais existant pour les hispanophones lors de la présentation audiovisuelle du contraste non-natif le plus visible, avec des scores de 87 % en modalité auditive et 88.6% lors de la présentation bimodale. Les scores des japonais augmentaient également de 1.3% entre les deux conditions, mais dans une moindre mesure, probablement dû au fait que cette population donne moins de poids aux informations visuelles. Une fois les données transformées afin d'éviter les biais dans les réponses, les différences entre les modalités auditive et audiovisuelle restaient significatives. Cependant, l'ajout de la modalité visuelle n'impactaient pas les résultats lors de l'identification des phonèmes peu visibles /r/ et /l/.

2.2.3.2 CARACTERISTIQUES VISUELLES ET ACOUSTIQUES DE LA LANGUE MATERNELLE

Le poids donné aux informations visuelles ou biais visuel est également modulé par les caractéristiques visuelles et acoustiques des phonèmes de la langue maternelle (i.e., le niveau d'informativité de chacun des canaux), car même dans le cadre de la perception des phonèmes natifs, certaines populations, comme les japonais (Sekiyama & Tohkura, 1991) ou les chinois (Hayashi & Sekiyama, 1998) ont une tendance à moins considérer les informations visuelles (pour les raisons de cette sous-utilisation des informations visuelles cf. Chapitre 2.1.5.2. « Facteurs cognitifs, langagiers et différences inter-individuelles ») par rapport aux américains par exemple. Cela serait dû à l'organisation de l'espace phonologique de la langue maternelle. Dans ce sens, citons Wang et al. (2009) qui obtenaient une amélioration des performances pour des coréens et des chinois parlant le mandarin lors de la présentation audiovisuelle de l'interdentale fricative non-native, avec une utilisation plus efficace des informations visuelles par les premiers. Les auteurs suggèrent que l'interdentale serait plus difficile pour les chinois car cette consonne a, dans leur répertoire phonologique, plus de voisin (i.e., de phonèmes

proches) que n'en ont les coréens. En effet, ces derniers ne possèdent que la fricative alvéolaire alors qu'en mandarin, l'alvéolaire et la labiodentale existent. Cela génère de l'ambigüité chez les participants chinois, ils auraient donc tendance à plus porter leur attention sur les informations visuelles sans pour autant pouvoir désambiguïser les visèmes. En effet, les coréens surpassent les chinois lors de la perception audiovisuelle, attestant d'une utilisation plus efficace des informations visuelles lorsqu'elles sont présentes avec le signal auditif. La saturation du répertoire phonologique sur des caractéristiques spécifiques modulerait donc l'utilisation des indices acoustique et visuel. Il semblerait également que plus on porte d'attention sur les visèmes dans sa langue maternelle, plus nous sommes susceptibles de les utiliser dans le cadre de la perception de al L2 (Hardison, 1999 ; Wang et al., 2008).

2.2.3.3 RELATION ENTRE LES VISEMES DE LA L1 ET DE LA L2

Des modulations sont également induites par les relations qui peuvent exister entre les visèmes de la L1 et de la L2. Ces relations ont été formulées par Hazan et al. (2006). Celles-ci suivraient les mêmes règles de similarité L1/L2 que la perception auditive phonèmes non-natifs des modèles PAM (Best & Tyler, 2007b ; Best, 1995) et SLM (Flege, 1995). Les indices visuels de la L2 peuvent être classés comme « identique », « similaire », ou « nouveaux », en fonction de l'existence ou non de leur contrepartie dans la L1. Dans le modèle de Hazan et al. (2006), trois types de relation sont décrites entre les catégories visuelles de parole entre L1 et L2. La première catégorie visuelle concerne les visèmes qui existent à la fois dans la L1 et la L2 et qui marque les mêmes distinctions phonémiques (e.g., /f-s/ en français et espagnol). Dans ce cas, l'utilisation des indices visuels ne posera pas de difficulté particulière. C'est ce qu'a montré notamment Kluge, Reis, Nobre-oliveira, et Bettoni-techio, (2006) et Kluge, (2009) qui se sont intéressés à l'utilisation des indices visuels permettant de distinguer /m/ et /n/ par des locuteurs du portugais brésilien apprenant l'anglais. Ces deux phonèmes bénéficient en anglais d'une différence dans leur place d'articulation (bilabiale vs alvéolaire), or le portugais parlé par les brésiliens n'a pas de distinction de place pour les nasales lorsqu'elles sont en position de coda (position finale de la syllabe). Celles-ci ne sont en effet pas complètement réalisées en position finale. Elle teste donc des visèmes (bilabiale et alvéolaire) qui existent donc dans la langue maternelle des participants mais dont les distinctions phonémiques ne sont pas utilisées après une voyelle en position finale. Un test d'identification était administré en modalité auditive, audiovisuelle et visuelle, les items

présentés étaient des séquences monosyllabiques CVC (e.g., tim/tin, gem/gen et cam/can). Les résultats des deux études montrent une amélioration des performances de 12.8 % pour le phonème /m/ et de 19,5% pour le phonème /n/ lors de la présentation audiovisuelle par rapport à la présentation auditive seule. Hazan et al. (2006) mettent également en évidence que les espagnols ont de meilleures performances que les japonais lors de la perception audiovisuelle du contraste /b/-/v/ car ils possèdent le visème de la fricative labiodentale /f/ alors qu'il n'existe pas en japonais.

La deuxième catégorie décrite dans le modèle renferme les visèmes qui existent dans la L2 mais pas dans la L1 (e.g., /θ/ anglais pour les français) ; et la dernière est une catégorie visuelle qui existe à la fois dans la L1 et dans la L2 mais qui est utilisée pour des distinctions phonétiques différente dans la L2 (e.g., le /v/ anglais qui est parfois réalisé en espagnol suite à une assimilation de voisement). Dans ces deux derniers cas, les individus ont possiblement perdu la sensibilité à ces indices qui ne sont pas utilisés dans la L1. Des études semblent confirmer le deuxième comme le troisième axes du modèle (Hazan et al., 2006, 2005 ; Ortega-Llebaria et al., 2001 ; Werker et al., 1992) alors que d'autres semblent les contredire. Wang et al., (2008, 2009) obtiennent une amélioration lors de la perception audiovisuelle du phonème /θ/ qui n'existe pas dans le répertoire visuel des participants chinois et coréens. Hardison (1999, 2003, 2005) observe quand à lui une amélioration lors de la perception du contraste /r/-/l/ perçu par des japonais.

2.2.3.4 LE POIDS RELATIF DES INFORMATIONS VISUELLES

Des études ont montré que les individus ont tendance à se reposer d'avantage sur les informations visuelles lorsque le locuteur s'exprime dans une langue étrangère, même quand les phonèmes utilisés existent dans la langue maternelle du participant. Cela est également le cas lorsqu'un compatriote s'exprime avec un accent régional (Irwin, Pilling, & Thomas, 2011). Cet effet est appelé le « *non-native talker effect* » ou « *foreign language effect* » (Sekiyama & Tohkura, 1993 ; Chen & Hazan, 2007 ; Fuster-Duran, 1996 ; Hazan & Li, 2008 ; Wang et al, 2008). Ces études, dont la plupart utilisent l'effet McGurk, ont en effet montré que lorsqu'un individu perçoit des stimuli non congruents émis dans une langue étrangère, il donne plus de réponses « visuelles » que lorsqu'il perçoit les mêmes stimuli enregistrés par un locuteur natif.

L'étude de Werker et al. (1992) a en effet révélé une relation inverse entre la maîtrise d'une langue et l'utilisation de l'information visuelle. Dans cette étude, lorsqu'un francophone canadien était face à un stimulus incongruent composé d'un /θ/ visuel (place d'articulation qui n'existe pas en français) et d'un /ba/ auditif, les participants francophones avaient tendance à se donner des réponses d'autant plus « visuelles » lorsqu'ils avaient peu de connaissance dans la langue. Cependant, le fait que les auditeurs se reposent plus sur l'information visuelle lorsque les séquences sont produites par un locuteur étranger ne va pas de pair avec une amélioration des performances. En effet, les participants francophones de leur étude ne reportaient pas avoir perçu une interdentale mais /da/ ou /ta/. Cependant, aucune condition audiovisuelle congruente n'a été inclue au design expérimental. Il nous est donc impossible de savoir comment sont utilisées ces informations dans le cadre de la perception de stimuli audiovisuels congruents. Même si la plupart des études atteste d'un effet positif de la présentation audiovisuelle, le gain se révèle en réalité être plus important dans le cadre de la perception de phonèmes natifs que non-natifs (Gelder & Vroomen, 1992 ; Hayashi & Sekiyama, 1998 ; Hazan et al., 2006 ; Ortega-Llebaria et al., 2001 ; Wang et al., 2008, 2009; Werker et al., 1992). En 2013, Yi, Phelps, Smiljanic et Chandrasekaran ont réalisé un set d'expérience qui vont en effet dans ce sens. Lors de parole perçue dans le bruit, le gain pour les natifs était plus grand que celui des percevants non natifs (Figure 14).

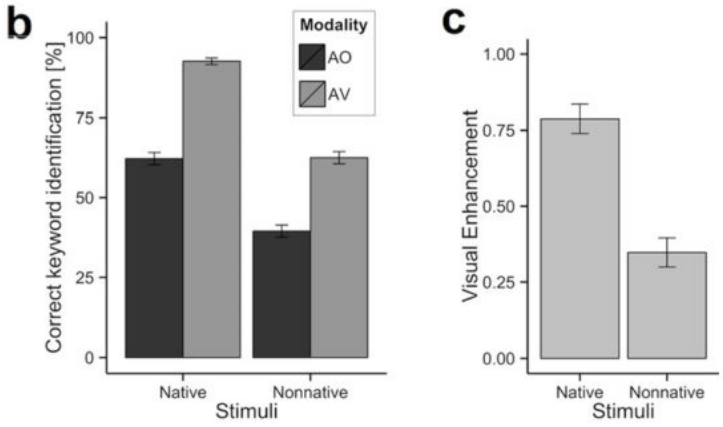

Figure 12. (b) Pourcentage de mots clés correctement identifies par des anglophones (Native) et des coréens (Nonnative), lors de présentation Auditive (AO) ou Audiovisuelle (AV); (c) Mesures d'amélioration fournie par les informations visuelles [(AV-AO)/(1-AO)] en fonction du groupe. (Extrait de Yi et al., 2013)

Il apparaît donc que malgré le fait que l'on porte plus d'attention aux informations visuelles quand le locuteur est étranger ne soit pas une raison suffisante pour utiliser indices visuels fournis par les mouvements articulatoires aussi efficacement que dans le cadre de la langue

maternelle. Ces résultats nous montrent que les variations de réalisation articulatoire entre la langue maternelle et la langue étrangère modulent l'efficacité avec laquelle nous utilisons des indices visuels qui ne nous sont pas familiers. Cependant, avec l'expérience, ces résultats tendent à changer. Wang et al. (2009) observent en effet une corrélation positive entre la longueur du séjour dans un pays étrangers et l'amélioration des performances liée à l'utilisation des informations visuelles. Donc plus ont devient expert dans une langue, plus l'information visuelle permet d'améliorer les performances, malgré le fait que l'on se repose moins sur ces dernières. Cela serait dû au fait qu'à plus ou moins long terme, les mouvements labiaux sont associés aux composantes phonologiques de la langue.

CHAPITRE 3

ETUDE 1

MODULATION DE LA SURDITE PHONOLOGIQUE LORS DE LA PERCEPTION AUDIOVISUELLE DE CONTRASTES NON NATIFS PAR DES MONOLINGUES FRANCOPHONES ET HISPANOPHONES

3.1 INTRODUCTION

De nombreuses études attestent du bénéfice lié à la perception audiovisuelle de la parole. En effet, cette présentation permet, dans le cadre de la langue maternelle, d'améliorer l'intelligibilité du discours (Grant & Seitz, 2000), mais elle améliore également les capacités d'identification de phonèmes ou de mots dans le bruit . Cet avantage est en majeure partie dû à la nature même du signal audiovisuel. En effet, les mouvements articulatoires visibles sont utilisables de manière précoce. Ils sont souvent présents et exploitables avant le signal acoustique (Chandrasekaran, Trubanova, Stillittano, Caplier et Ghazanfar, 2009 ; Schwartz et Savariaux, 2013 pour des études en production ; Cathiard, 1994 ; Escudier, Benoit et Lallouache, 1990 ; Jesse & Massaro, 2010; Smeele, 1994 pour des études en production), même si l'inverse peut être observé sous certaines conditions ou pour certains types de stimuli (Schwartz & Savariaux, 2013 ; Troille et al., 2007, 2010). Cette caractéristique du signal audiovisuel de parole permet d'une part de réduire l'ambiguïté des entrées auditives, mais également de prédire certaines caractéristiques du signal acoustique qui va suivre (Paris et al., 2013). Enfin, la prise en compte des deux informations relève de processus automatiques comme l'atteste l'effet McGurk (McGurk & MacDonald, 1976).

Cependant ces études se focalisent sur la langue maternelle, laissant en suspens la question de l'utilisation de ces informations dans le cadre de la perception de phonèmes étrangers. En effet, la question sous-jacente serait de savoir si ces informations sont utilisables dans le cas de la perception de phonèmes non natifs, et ainsi, d'envisager la possibilité que ces indices visuels puissent améliorer les capacités de discrimination de phonèmes non-natifs, et donc avoir un effet sur la surdité phonologique. Dans le cas des apprenants, la question a été soulevée, cependant les résultats de la littérature restent très hétérogènes. Alors que certaines études ont montré un impact des indices visuels pour l'identification des contrastes non-natifs chez les apprenants (Davis & Kim, 2004 ; Erdener & Burnham, 2005 ; Hardison, 1999, 2003, 2005 ; Hayashi & Sekiyama, 1998 ; Hazan et al., 2006 ; Kluge et al., 2006 ; Sekiyama, Kanno, Miura, & Sugita, 2003 ; Thompson & Hazan, 2010 ; Wang et al., 2008, 2009), d'autres n'ont pas mis à jour d'avantage (Hazan et al., 2002, 2005 ; Ortega-Llebaria, Faulkner, & Hazan, 2001). Les informations visuelles seraient donc utilisables sous certaines conditions. L'étude de Hazan et al. (2006) montre par exemple que si le visème est partagé dans les deux langues des participants, un avantage peut être observé. C'est par exemple le cas du visème /f/ qui est connu par les hispanophones et qui peut donc être utilisé pour

discriminer le contraste /b/-/v/ (qui n'existe pas en espagnol). De plus, si les articulateurs recrutés lors de la production ne sont pas assez visibles, aucun avantage ne sera obtenu (Hazan et al., 2006, 2005 ; Wang et al., 2008 ; voir Hardison, 2003 pour des résultats contradictoires), comme c'est le cas pour la langue maternelle (Smeele, 1994). Mais le poids relatif donné à chacun des canaux, ainsi que la structure même du répertoire phonologique natif module l'utilisation des informations visuelles (Wang et al., 2009).

La question que nous posons ici est de savoir si les informations visuelles pourraient être utilisées même sans connaissance préalable dans une langue. En effet, cette capacité à discriminer des phonèmes dès le début de l'apprentissage est cruciale puisque si l'identification d'un phonème échoue, ce sont toutes les étapes successives de traitement de la parole qui seront biaisées. Cependant, très peu d'études se sont tournées vers les capacités perceptives des individus sans connaissances préalables dans une langue étrangère. Si l'information articulatoire est utilisée de manière automatique, est-ce le cas lorsque nous sommes confrontés à des visèmes qui sont inconnus de l'observateur ? Les études de Davis et Kim (1998, 1999, 2001) indiquent que les informations visuelles permettent à la fois de mieux produire des phonèmes non natifs, mais également de mieux les percevoir.

Au regard des études mentionnées ci-dessus, nous nous interrogeons sur les capacités de discrimination de contrastes phonologiques n'existant pas dans la langue d'individus monolingues, mais également sur les capacités de discrimination de ces mêmes contrastes par des individus parlant cette langue. L'objectif de cette étude est d'évaluer les capacités perceptives de monolingues francophones et hispanophones lors de la perception d'un contraste phonologique qui existe en espagnol mais qui n'existe pas en français et vice-versa. Cette étude est donc composée de deux expériences dans lesquelles la tâche du participant (i.e., discrimination de phonèmes) et le protocole expérimental étaient les mêmes afin de pouvoir généraliser les résultats d'une population à l'autre. Les contrastes étudiés sont /b/-/v/ et /f/-/θ/. Le premier contraste existe en français. Les phonèmes /b/ et /v/ sont en effet utilisés pour distinguer par exemple les mots « beau » et « veau ». Ce contraste n'existe pas en espagnol car /v/ ne fait pas partie des phonèmes de l'espagnol. Le second contraste (/f/-/θ/) existe en espagnol mais pas en français puisque /θ/ ne permet pas de constituer un mot de la langue française. Ainsi, lors de la perception de /θ/ un francophone aura tendance à l'assimiler à un des phonèmes qui existent dans son répertoire phonologique (Berger, 1951; Brannen, 2002). Les deux groupes de participants, espagnols et français, seront testés lors de deux

expériences où leurs capacités de discrimination des deux contrastes seront évaluées. Le but de l'étude est dans un premier temps de mettre à jour une surdité phonologique lors de la perception *auditive* d'un phonème non natif. Cela se traduirait, chez les francophones, par des difficultés à discriminer /f/ et /θ/, puisque le phonème non natif devrait être assimilé à une catégorie phonologique native la plus proche, en l'occurrence /f/ puisque ces phonèmes sont contigus sur le continuum des fricatives non voisées. En effet, selon le modèle PAM (*Perceptual Assimilation Model*, Best et al., 2001, 2003, 1988, 2007a ; Best, 1994, 1995), le phonème non-natif (i.e., /θ/) devrait être assimilé à la catégorie native /f/ ce qui donnerait lieu à une surdité phonologique. De la même façon, chez les hispanophones, /v/ devrait être assimilé à /b/ et induire une surdité phonologique. La surdité devrait donc se manifester pour les deux groupes par des scores de discrimination faible ainsi que des temps de réaction plus importants lors de la présentation du phonème non natif mais de bonnes performances lors de la perception de phonème natif. Nous supposons également que les réponses fournies sur le phonème non natif devraient être plus rapides lorsque le phonème est perçu par le groupe d'individus natifs que par les non natifs.

Par contraste, lors de la présentation audiovisuelle, nous pensons que cette surdité pourra être réduite notamment via l'utilisation des informations visuelles fournies par les mouvements articulatoires du locuteur. Les deux phonèmes d'intérêt ont d'ailleurs été sélectionnés pour leur place d'articulation saillante, puisque /θ/ est une fricative interdentale qui nécessite de placer la langue entre les dents. Ce phonème est l'un des phonèmes les plus faciles à identifier visuellement (Mills, 1987). Le phonème de /v/ est également visuellement facilement distinguable de /b/ puisque la première consonne est une labiodentale alors que la seconde est bilabiale. Ces différences articulatoires sont très visibles et devraient permettre de limiter l'effet de la surdité phonologique. Si les informations visuelles sont utilisées par les participants, cela devrait se traduire par une diminution du phénomène de surdité phonologique.

Enfin, l'observation des performances obtenues par les groupes contrôles nous permettront d'évaluer l'apport des informations visuelles et de savoir si celles-ci sont suffisantes pour ramener une qualité de perception équivalente à celle des natifs. Pour cela, chacune des populations servira de groupe contrôle lors de la perception du contraste natif et de groupe expérimental lors de la perception du contraste non natif.

3.2 MATERIEL ET METHODES

3.2.1 PARTICIPANTS

Les deux groupes étaient respectivement composés de 20 monolingues francophones (*M* = 22,1 ans ; *SD* = 2.9) testés à l'université de Grenoble et 20 monolingues hispanophones (*M* = 25,5 ans ; *SD* = 3.47) testés au Center of Brain and Cognition de Barcelone, Universitat Pompeu Fabra, Espagne. Les participants remplissaient un questionnaire (Marian, et al, 2007) sur les langues apprises/pratiquées ainsi que sur leurs éventuels voyages à l'étranger (Annexe 1).

3.2.2 MATERIEL

3.2.2.1 CONTRASTE /f/-/θ/

Quinze enregistrements de 18 séquences bi-syllabiques (i.e., CVCV) et quatre séquences monosyllabiques (i.e., CV) prononcées par un locuteur hispanophone ont été enregistrés. Sur les vidéos et photographies utilisées durant l'expérience, le visage complet du locuteur apparaît. Seulement 11 réalisations de quatre monosyllabes (i.e., /fa/, /fe/, /θa/, /θe/) et six bisyllabes (i.e., /pafo/, / pefo /, /paθo/, /peθo/, /peso/, /pafo/) ont été conservées en fonction de la qualité des réalisations (e.g., évitement de clignement des yeux ou des bruits de langue) pour un total de 110 stimuli qui furent par la suite segmentés manuellement à partir de la première image d'ouverture des lèvres jusqu'à la première image de fermeture labiale. Les stimuli ont été présentés en modalité auditive (le visage du locuteur apparaissait à l'écran) et audiovisuelle dans des blocs contrebalancés par participant. A l'intérieur de chaque bloc, trois listes variant sur la structure des stimuli (i.e., CV*1 et CVCV*2) étaient présentées. Le contexte vocalique pouvait également varier. Pour les séquences CV, le contexte était toujours /e/ (e.g., /fe/) ; pour les séquences CVCV, plusieurs configurations vocaliques étaient possibles : -a-o, -e-o. Chaque liste était composée (1) de séquences qui contenaient le phonème *natif* (N), qui existe en français : /f/ (i.e., /fe/, /pafo/ et /pefo/) ; (2) et de séquences qui contenaient un phonème *non natif* (NN), qui n'existe pas en français : /θ/ (i.e., /θe/, /paθo/ et /peθo/) ; et (3) de deux types de distracteurs qui variaient soit sur la consonne cible (SD1), soit sur les voyelles précédant ou succédant la consonne cible (SD2) selon le ratio 4 :2 :1 :1 (i.e., 28N, 14 NN, 7 SD1 et 7 SD2 dans chaque liste). Par exemple, si le contraste /f/-/θ/ était

testé dans un contexte vocalique « -a-o », la liste était composée de 28 /pa**f**o/, 14 /pa**θ**o/, 7 /p**e**so/ et 7 /pa**f**o/. Les stimuli distracteurs étaient présentés afin de donner du sens à la tâche. En effet, en supposant que la surdité phonologique soit importante pour les participants francophones, les participants ne feront pas de différence entre les séquences contenant le phonème natif et le phonème non natif. Si c'est le cas, ils se retrouveront face à une tâche où ils répondront constamment que les stimuli sont les mêmes. Afin d'éviter ce problème nous avons inclus des distracteurs dans un ratio suffisant pour limiter le désengagement du participant. Pour chaque type de stimuli (i.e., N, NN et SD), chaque exemplaire était aléatoirement sélectionné par le logiciel Eprime (e.g., si SD1 était /pafo/, les exemplaires présentés pouvaient être "pafo1-pafo6-pafo11-pafo5-pafo10-pafo8-pafo12". L'ordre de présentation des stimuli dans la liste était randomisé. Un total de 168 stimuli était présenté dans chaque condition (i.e., auditive et audiovisuelle).

3.2.2.2 CONTRASTE /b/-/v/

Quinze enregistrements de huit séquences bisyllabiques (CVCV) et quatre séquences monosyllabiques (CV) prononcées par un locuteur français natif ont été enregistrées. Les fichiers obtenus ont été soumis aux mêmes traitements que ceux précédemment cités. Onze productions de 4 séquences monosyllabiques (i.e., /pabo/, /pebo/, /pavo/, /pabo/, /paro/, /pero/) et 6 bisyllabiques (i.e., /be/, /ve/, /bo/, /vo/, /re/, /ro/) ont été sélectionnées sur la base de la qualité de la réalisation pour un total de 110 stimuli. Le contexte vocalique était le même que précédemment. Chaque liste était composée (1) de séquences qui contenaient le phonème natif (N), qui existe en espagnol : /b/ (i.e., /be/, /pabo/ et /pebo/) ; (2) de séquences qui contenaient un phonèmes non natif (NN), qui n'existe pas en espagnol : /v/ (i.e., /ve/, /pavo/ et /pevo/) ; et (3) de deux types de distracteurs qui variaient soit sur la consonne cible (SD1), soit sur les voyelles précédant ou succédant la consonne cible (SD2) selon le même ratio que l'expérience précédente. La suite du matériel a été construite en suivant les mêmes principes que ceux du contraste /f/-/θ/.

3.2.3 PROCEDURE

Les participants ont été testés seuls dans une chambre sourde. Les participants francophones et hispanophones étaient testés sur les deux contrastes, de manière successive. L'ordre des expériences était contrebalancé. Chaque liste se déroulait en deux temps. Le participant était d'abord exposé à un *« mot » de référence* (contenant le phonème natif) qui consistait en six présentations de différents exemplaires du même stimulus de façon à ce qu'il mémorise phonologiquement cette référence (Figure 15). Le terme « mot espagnol » était utilisé dans la consigne afin de limiter une mémorisation basée sur les propriétés acoustiques du signal et maximiser l'encodage phonologique. Il devait l'écouter lors des blocs auditifs ou écouter et regarder le locuteur prononcer cette séquence en condition audiovisuelle. Ensuite, le participant était exposé à une « liste complète » qui contenait les stimuli N, NN et SD. A chaque stimulus présenté, le participant devait répondre le plus rapidement possible (même pendant celui-ci) pour indiquer s'il pensait que le « mot » entendu était le même ou un « mot » différent de son mot de référence.

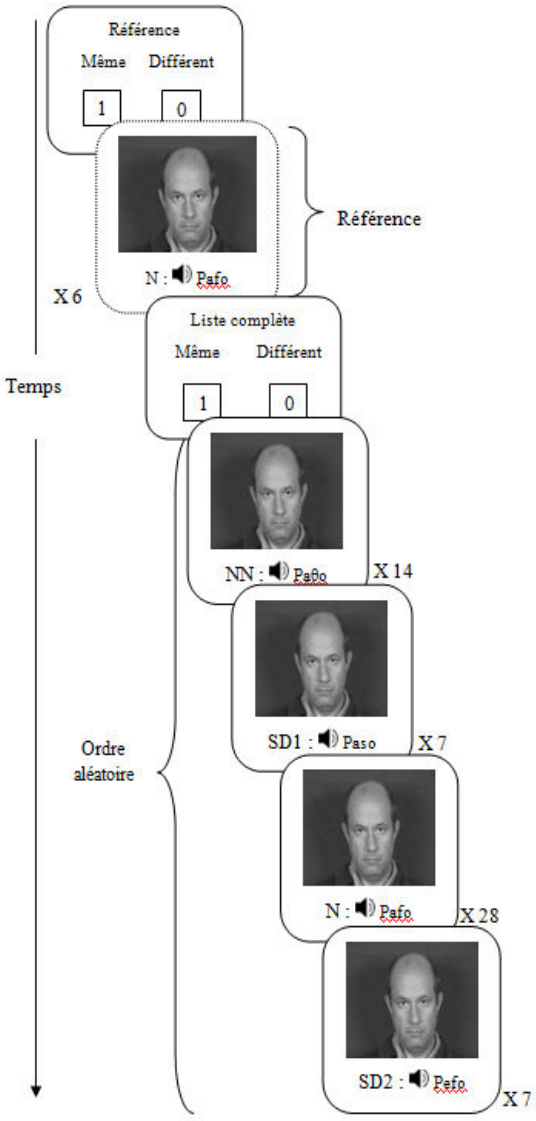

Figure 13. Représentation schématique de la procédure expérimentale.

Une session d'entraînement en modalité audiovisuelle fut réalisée avant le test. Les stimuli présentés n'étaient pas les mêmes que ceux présentés durant l'expérience. La procédure était la même que celle de la condition expérimentale mais ne contenait que 19 stimuli au même ratio que ceux précédemment décrits dans la section « Matériel ».

Les participants francophones étaient donc exposés à 4 blocs : (1) stimuli espagnols en auditif, (2), stimuli espagnols en audiovisuel, (3) stimuli français (i.e., contraste /v-b/ en auditif et (4) stimuli français en audiovisuel. Tous ces blocs étaient contrebalancés par participants.

3.3 RESULTATS

Les mêmes analyses ont été réalisées pour les deux expériences. Deux analyses de la variance (ANOVA) sur la variable Groupe (test ; contrôle) en facteur inter-participant et deux facteurs intra-participant : Modalité (auditive ; audiovisuelle) et Type d'item (N ; NN) ont été réalisées sur le Taux de réponses correctes et le Temps de réaction pour les réponses correctes. Les temps de réaction supérieurs à 2500 ms et inférieurs à 250 ms ont été supprimés (environ 1% des essais pour le contraste /f/-/θ/ et 0.5% des essais pour le contraste /b/-/v/).

3.3.1 CONTRASTE /f/-/θ/

3.3.1.1 TAUX DE REPONSES CORRECTES

L'ANOVA a révélé un effet principal du Groupe ($F_{(1,38)} = 30.08$, $p < .0001$) avec des scores plus importants pour les hispanophones, qui obtiennent 89% de réponses correctes ($SD = 2$) contre seulement 78% pour les francophones ($SD = 5$). L'effet de la Modalité de présentation est également significatif, $F_{(1,38)} = 126.15$, $p < .001$. Les participants sont plus précis lorsqu'ils ont accès aux deux informations ($M = 92\%$; $SD = 2$) que lorsqu'ils n'ont accès qu'aux informations auditives ($M = 75\%$; $SD = 4$). Nous observons également un effet du Type d'item ($F_{(1,38)} = 43.03$, $p < .001$), avec un pourcentage de réponses correctes plus important lorsque le phonème natif est présenté ($M = 90\%$; $SD = 1$) par rapport au phonème non natif ($M = 77\%$; $SD = 4$). Les interactions Type d'item X Groupe ($F_{(1,38)} = 8.049$, $p < .001$) et Modalité X Type d'item ($F_{(1,38)} = 51.10$, $p < .001$), sont significatives ainsi que l'interaction triple des facteurs Groupe X Modalité X Type d'item ($F_{(1,38)} = 10.70$, $p < .001$). Les résultats sont présentés dans la Figure 16.

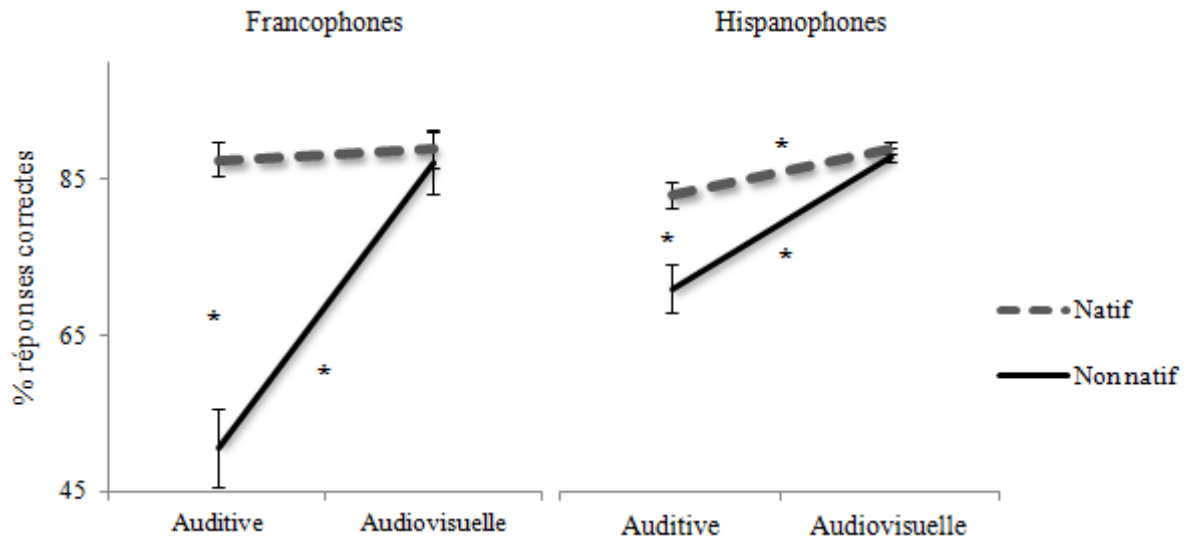

Figure 14. Pourcentage de réponses correctes moyen en fonction du Groupe (test ; contrôle, du Type de stimuli (N /f/ ; NN /θ/) et de la Modalité de présentation (auditive ; audiovisuelle).

* : $p < .05$

Les comparaisons planifiées réalisées sur le groupe test ont révélé un phénomène de surdité phonologique en présentation auditive ($F_{(1,38)} = 71.69$, $p < .001$) avec un taux de réponses correctes de 50.7% ($SD = 5$) lorsque des séquences contenant le phonème qui n'existe pas dans la langue sont présentées contre 87,5% ($SD = 2$) quand elles contiennent le phonème qui existe dans la langue. Cette différence disparait en présentation audiovisuelle ($F < 1$) car les réponses correctes sur le phonème qui n'existe pas en français augmente avec la présentation audiovisuelle ($F_{(1,38)} = 81.54$, $p < .001$) notamment grâce à l'information supplémentaire fournie par les mouvements labiaux, saillants pour ce contraste.

En ce qui concerne les résultats du groupe contrôle, on constate que leur taux de réponses correctes est plus important que celui des francophones de façon générale durant la présentation auditive ($F_{(1,38)} = 23.47$, $p < .001$) et audiovisuelle ($F_{(1,38)} = 15.80$, $p < .001$). Même si le groupe contrôle obtient des performances différentes entre les stimuli natifs et les non natifs en modalité auditive ($F_{(1,38)} = 10.60$, $p < .01$), leur précision est de 75.4% ($SD = 3$) contre seulement 50.7% pour les francophones ($F_{(1,38)} = 21.96$, $p < .001$). Comme pour les francophones, les différences obtenues pour les stimuli natifs et les non natifs disparaissent en présentation audiovisuelle ($F<1$). Le groupe contrôle bénéficie d'une aide de la présentation audiovisuelle sur la natifs ($F_{(1,38)} = 24.87$, $p < .001$) ainsi que pour les stimuli non natifs ($F_{(1,38)} = 14.49$, $p < .001$).

3.3.1.2 TEMPS DE REACTION

L'ANOVA a révélé un effet principal du Groupe ($F_{(1,38)} = 5.13$, $p < .05$), avec des temps de réaction plus importants pour les hispanophones (1082 ms ($SD = 45$) contre 990 ms ($SD = 39$) pour les francophones). L'effet de la Modalité est également significatif ($F_{(1,38)} = 49.17$, $p < .001$) avec des temps de réaction moins importants lors de la présentation audiovisuelle ($M = 972$ ms ; $SD = 48$) par rapport à la présentation auditive ($M = 1099$ ms ; $SD = 41$). Le Type d'item modulent également les temps de réaction, $F_{(1,38)} = 79.75$, $p < .001$). Les participants mettent globalement plus de temps pour répondre quand la séquence contient un phonème non natif ($M = 1083$ms ; $SD = 21$) que lors qu'elle contient un phonème natif ($M = 989$ ms ; $SD = 16$). L'interaction Modalité X Type d'item ($F_{(1,38)} = 9.08$, $p < .005$) est significative. Les résultats sont présentés dans la Figure 17.

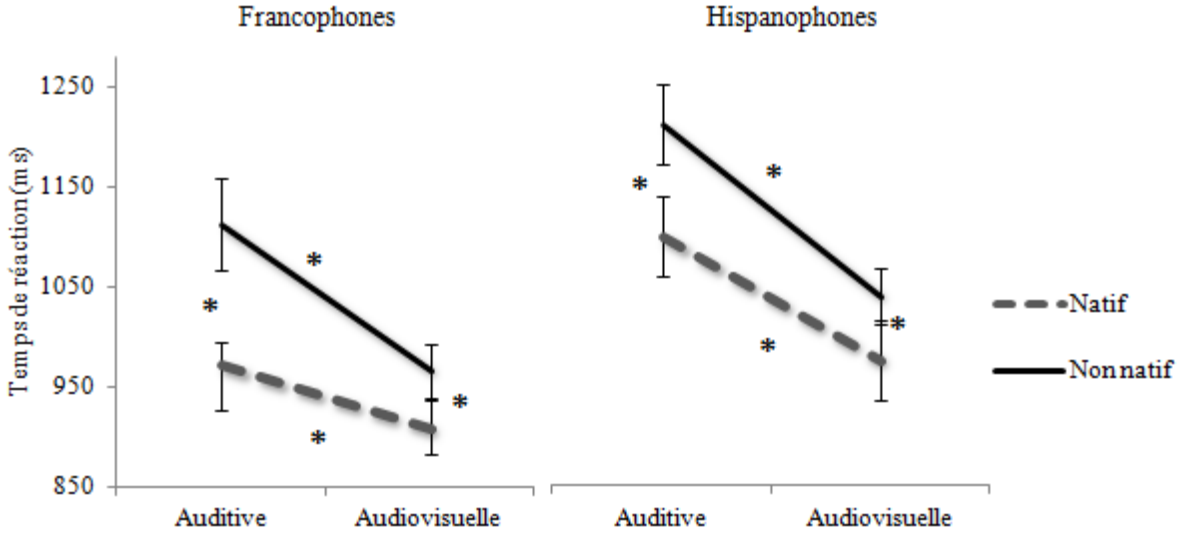

Figure 15. Temps de réaction (ms) moyen en fonction du Groupe (test ; control), du Type de stimuli (N /f/ ; NN /θ/) et de la Modalité de présentation (auditive ; audiovisuelle).
* : $p<.05$

Malgré l'absence d'interaction triple entre nos trois facteurs d'intérêts, nous allons tout de même opérer les mêmes comparaisons planifiées que celles effectuées sur le taux de réponses correctes car nos hypothèses principales portent sur l'effet de la Modalité de présentation et du Type de phonème en fonction du Groupe sur les temps de réponses. Celles-ci ont révélé un coût lié à la présentation auditive des phonèmes qui n'existent pas dans la langue maternelle par rapport au phonème natif ($F_{(1,38)} = 43.23$, $p < .001$) et cette différence persiste lors de la présentation audiovisuelle où les temps de réaction pour le phonème natif sont plus courts que lors de la présentation du phonème non natif ($F_{(1,38)} = 6.91$, $p < .05$). Ce

coût de traitement est cependant atténué lors de la présentation audiovisuelle du phonème natif, $F_{(1,38)} = 150.20$, $p < .005$. Cependant on constate que même lorsque le phonème présenté n'existe pas dans la langue maternelle, celui-ci profite tout de même de la présentation audiovisuelle car le temps de réponse diminue par rapport à une présentation auditive, $F_{(1,38)} = 15.56$, $p < .005$. Le même pattern se retrouve chez les contrôles, avec une accélération du traitement liée à la présentation audiovisuelle, que ce soit sur les phonèmes qui existe dans la langue maternelle ($F_{(1,38)} = 39.24$, $p < .001$) ou sur les phonèmes non natifs ($F_{(1,38)} = 21.35$, $p < .001$). On remarque également pour ce groupe que les temps de réaction sont globalement plus importants lors de la présentation auditive ($F_{(1,38)} = 5.33$, $p < .05$). La différence lors de la présentation audiovisuelle n'étant que tendancielle ($F_{(1,38)} = 3.25$, $p = .079$).

3.3.2 CONTRASTE /b/-/v/

3.3.2.1 POURCENTAGE DE REPONSES CORRECTES

L'ANOVA n'a révélé aucun effet principal, que ce soit pour le Groupe ($F < 1$), la Modalité ($F_{(1,38)} = 1.78$, $p = .19$), ou le Type d'item ($F < 1$) . Les résultats sont présentés dans la Figure 18.

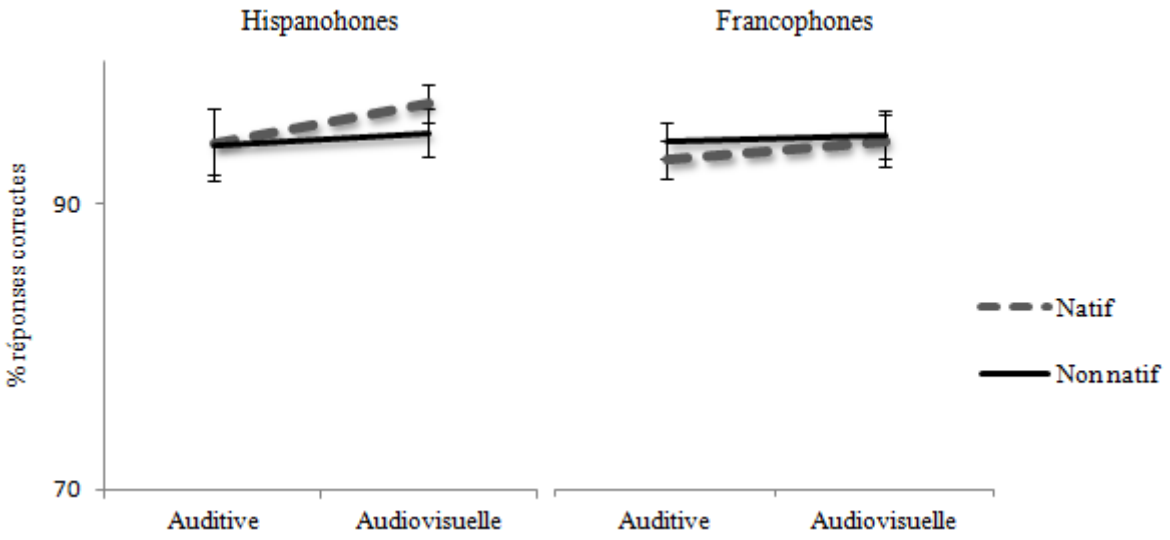

Figure 16. Pourcentage de réponses correctes moyen en fonction du Groupe (test ; contrôle), du Type de stimuli (N /b/ ; NN /v/) et de la Modalité de présentation (auditive ; audiovisuelle).
* : $p < .05$

3.3.2.2 TEMPS DE REACTION

L'ANOVA a révélé un effet principal du Groupe ($F_{(1,38)} = 18.11$, $p < .001$) avec un avantage du groupe contrôle de francophones ($M_{contrôle} = 893$ ms, $SD = 26$; $M_{test} = 1029$ ms, $SD = 38$). La Modalité de présentation module également les temps de réponses, $F_{(1,38)} = 11.06$, $p < .01$. Ceux-ci sont moins importants lors de la présentation bimodale ($M = 939$ ms, $SD = 37$) qu'unimodale ($M = 982$ ms, $SD = 30$). L'effet du Type d'item est également significatif ($F_{(1,38)} = 64.371$, $p < .001$) avec des temps de réaction plus importants lors de la présentation de séquences contenant le phonème non natif ($M = 992$ ms ; $SD = 17$) que natif ($M = 930$; $SD = 17$). L'interaction Modalité X Type d'item est significative, $F_{(1,38)} = 6.43$, $p < .05$. L'interaction triple entre nos facteurs d'intérêt (Modalité X Type de phonème X Groupe) n'est pas significative, $F_{(1,38)} = 1.30$, $p = .25$. Les résultats sont présentés dans la Figure 19.

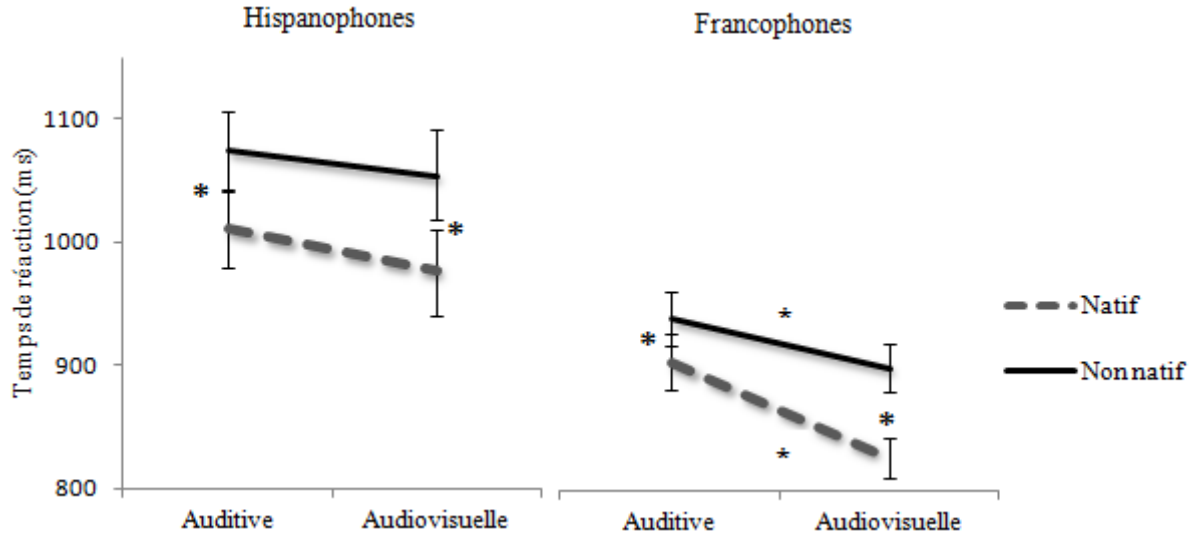

Figure 17. Temps de réaction (ms) moyen en fonction du Groupe (test ; control), du Type de stimuli (N /b/ ; NN /v/) et de la Modalité de présentation (auditive ; audiovisuelle).
* : $p < .05$

Compte tenu de nos hypothèses, les comparaisons planifiées ont été effectuées en considérant les trois facteurs. Pour le groupe test, les comparaisons planifiées ont mis à jour une absence d'apport de la modalité audiovisuelle, en effet les temps de réaction restent stables, pour les stimuli natifs ($F_{(1,38)} = 3.065$, $p = .08$), malgré une tendance à augmenter, ainsi que pour les stimuli non natifs ($F < 1$). L'avantage audiovisuel est par contre présent pour le groupe contrôle avec des réponses plus rapides par rapport à la présentation auditive pour les stimuli natifs ($F_{(1,38)} = 16.07$, $p < .001$) et les stimuli non natifs ($F_{(1,38)} = 4.12$, $p < .05$). Les différences entre ces deux types de stimuli sont significatives pour les

hispanophones lors de la présentation auditive ($F_{(1,38)}$ = 26.82, p < .001) et audiovisuelle ($F_{(1,38)}$ = 31.11, p < .001) ainsi que chez les francophones lors de la présentation auditive ($F_{(1,38)}$ = 8.28, p < .01) et audiovisuelle ($F_{(1,38)}$ = 27.26, p < .001).

3.4 DISCUSSION

Cette étude avait pour but de comparer les performances de discrimination lors de la perception de contrastes non natifs présentés en modalité auditive et audiovisuelle. Le contraste espagnol /f/-/θ/ ainsi que le contraste français /b/-/v/ ont été testés. Les participants étaient des francophones et des hispanophones monolingues évalués respectivement sur les deux contrastes afin que les données des deux groupes puissent à la fois être utilisées comme données test et contrôle. Les participants effectuaient une tâche de discrimination de phonèmes. Après avoir mémorisé un « mot » de référence, les participants devaient indiquer à chaque essai si le « mot » qu'ils percevaient était le même ou différent du « mot de référence ». Nous avons analysé le pourcentage de réponses correctes ainsi que les temps de réaction.

L'étude avait deux buts principaux. D'une part, il s'agissait de mettre en avant un phénomène de surdité phonologique lors de la perception auditive du phonème non natif. Ce phénomène a en effet été observé chez les francophones lorsqu'ils percevaient le phonème /θ/, mais pas chez les hispanophones percevant /v/. D'autre part, nous voulions étudier l'impact de la présentation audiovisuelle sur les capacités de discrimination des contrastes non-natifs. Pour les francophones, les informations labiales fournies lors de l'articulation de /θ/ étaient suffisamment saillantes et discriminantes pour améliorer les scores de 36.8%. Pour les hispanophones, qui ne subissaient pas de surdité phonologique, aucune modulation n'est observée, ni au niveau des pourcentages de réponses correctes, ni des temps de réaction. Cependant, ceux-ci présentaient des temps de réaction plus importants que les participants francophones quelque soit le contraste considéré.

3.4.1 CONTRASTE /f/-/θ/

Le groupe de francophones se comporte en accord avec nos hypothèses. En effet, nous retrouvons bien une surdité phonologique en modalité auditive sur le taux de réponses correctes. L'analyse des temps de réaction révèle également une difficulté de traitement des

stimuli non natifs puisque les temps de réaction sont systématiquement plus importants lors de la présentation du phonème non natif quelle que soit la modalité de présentation. Une différence significative existe entre les phonèmes natifs et les phonèmes non natifs, en faveur du premier type de stimulus. Cela traduit bien une difficulté lors de la perception des stimuli non natifs due au fait que les participants n'arrivent pas à interpréter correctement les indices auditifs ou même visuels fournis par le phonème /θ/.

On constate cependant qu'au niveau des réponses correctes, la présentation audiovisuelle améliore largement la perception des phonèmes inconnus. Nous obtenons en effet un taux de réponses correctes similaire à celui obtenu pour le phonème qui existe dans la langue maternelle (87,5% pour les non natifs contre 88,8% pour les natifs). De plus, les temps de réponse sont diminués lors de l'ajout d'informations visuelles (même s'ils restent supérieurs à ceux obtenus lors de la présentation audiovisuelle du phonème natif). Cette diminution du temps de réaction atteste d'une accélération des traitements due aux informations visuelles même lorsqu'on ne connaît pas le phonème présenté. Il semblerait donc que les mouvements articulatoires soient exploités même sans connaissances préalables du phonème perçu ou même du visème qui lui est associé.

Cela va à l'encontre des résultats de Hazan et al. (2006). Son modèle propose qu'un visème ne peut être utilisé pour désambigüiser des phonèmes que lorsqu'il fait partie du répertoire visèmique natif. Elle postule trois types de relations entre les visèmes de la L1 et de la L2. La première catégorie visuelle concerne les visèmes qui existent à la fois dans la L1 et la L2 et qui marquent les mêmes distinctions phonémiques (e.g., /f-s/ en français et espagnol). Dans ce cas, l'utilisation des indices visuels ne posera pas de différence particulière. La deuxième catégorie décrite dans le modèle renferme les visèmes qui existent dans la L2 mais pas dans la L1 (e.g., /θ/ anglais pour les français) ; et la dernière est une catégorie visuelle qui existe à la fois dans la L1 et dans la L2 mais qui est utilisée pour des distinctions phonétiques différentes dans la L2 (e.g., le /v/ anglais qui est parfois réalisé en espagnol suite à une assimilation de voisement). Dans ces deux derniers cas, les individus ont possiblement perdu la sensibilité à ces indices qui ne sont pas utilisés dans la L1 (Hazan et al., 2006, 2005 ; Ortega-Llebaria et al., 2001 ; Werker et al., 1992). Rappelons néanmoins que ces constats sont généralement observés sur des populations d'apprenants et non d'individus monolingues sans connaissances préalables des phonèmes présentés. Il se peut que cette variation concernant les populations étudiées induisent des différences dans les capacités à exploiter les mouvements oro-faciaux pour désambigüiser les phonèmes. Nos résultats seraient en accord avec ceux de

Werker et al. (1992). Ils ont en effet montré une relation inverse entre la maîtrise d'une langue et l'utilisation de l'information visuelle. Dans cette étude, lorsqu'un francophone canadien était face à un stimulus incongruent composé d'un /θ/ visuel (place d'articulation qui n'existe pas en français) et d'un /ba/ auditif, les participants francophones avaient tendance à donner des réponses d'autant plus « visuelles » lorsqu'ils avaient peu de connaissance dans la langue.

Le groupe contrôle a quant à lui un taux de réponses correctes plus important comme cela était attendu ainsi que des temps de réaction moins importants lors de la présentation audiovisuelle qu'en condition auditive. Ce dernier point atteste d'une facilitation liée à la présence des informations visuelles qui permettent d'accélérer le traitement des phonèmes. Cependant, un point module nos interprétations : la différence significative qui existe entre les stimuli natifs et non natifs pour le groupe contrôle. Cette différence entre les phonèmes natifs et les non natifs pour les hispanophones serait donc plus probablement due à la similarité acoustique des deux phonèmes. En effet, nous savons que certains sons de langage sont difficilement discriminables en modalité auditive. En français, c'est par exemple le cas de /m-n/, qui sont toutes deux des consonnes nasales, ce qui les rends peux distinguables du point de vue acoustique mais très différentes d'un point de vue visuel. De plus alors que les différents modes d'articulation sont facilement discriminables sur la base des seuls indices auditifs, les distinctions de place d'articulation sont quant à elles visibles. Or, ce contraste diffère sur la place d'articulation (labiodentale et interdentale). Best & McRoberts (2003) ont réalisé une étude qui avait au départ pour but de montrer que le déclin des capacités de discrimination auditive de contrastes non natifs dépendait des articulateurs impliqués lors de la production du phonème. Ils pensent en effet observer une absence de difficultés lorsque le contraste non-natif testé (i.e., bilabiale et alvéolaire éjective non voisés /p′–t′/) implique différents organes articulatoires (*between-organ contrast*) et une difficulté à discriminer les phonèmes lorsque les mêmes contraintes articulatoires étaient imposées à un contraste non natif (i.e., latérale fricatives voisées et non voisées /ɬ– ɮ/) ou même natif (i.e /s-z/) (*within-organ contrast*). Ils ont montré que même dans le cadre de la langue maternelle, certains contrastes *within-organ* peuvent être difficiles à percevoir. Il y a donc un effet non négligeable de l'impact des articulateurs pour la catégorisation des sons non-natifs, mais également natifs. Ces résultats ont été répliqués sur d'autres phonèmes (Kuhl et al., 2006). Cependant, la fricative inter-dentale et la labiodentale, même si leur articulation est proche, ne peuvent pas à proprement parler être considérées comme un contraste *within-organ* puisque leur production recrute différents articulateurs. Ce contraste est d'ailleurs décrit comme étant celui qui induit le plus

de confusions en anglais (Miller & Nicely, 1955). La source de cette confusion est la similarité des propriétés acoustiques puisque des mesures effectuées sur la friction ont montré que ces deux consonnes se caractérisaient par une distribution uniforme de l'énergie spectrale (Heinz & Stevens, 1961 ; Klatt, 1986)

Enfin, les hispanophones auraient dû, selon nos attentes, répondre plus rapidement que les monolingues en présentation auditive, comme en audiovisuel. Néanmoins, ceux-ci sont en moyenne supérieurs à ceux des francophones, ce qui est surprenant. En effet, percevant des phonèmes natifs, ce groupe est expert dans la perception des phonèmes présentés et devrait donc les traiter plus rapidement. Ce résultat sera interprété au regard des résultats de l'étude 2, dans la partie discussion.

3.4.2 CONTRASTE /b/-/v/

Les participants hispanophones ne semblent pas être sujets à la surdité phonologique pour le contraste /b/-/v/. Ces résultats sont partiellement en accord avec ceux de Hazan et al. (2006). En effet, leurs résultats montraient des capacités de discrimination préservées chez les hispanophones lors de la perception du contraste /b/-/v/ avec des pourcentages d'identification de 87% lors de la présentation auditive et 88.6% lors de la présentation audiovisuelle. Cependant, ces résultats étaient obtenus pour une population d'hispanophones apprenant l'anglais ce qui limite les comparaisons entre nos résultats et les leurs. En effet, l'expertise acquise par les apprenants peut expliquer à elle seule les facilités de discrimination obtenues dans ces études.

Nos résultats semblent contraster largement avec ceux de Pons, Lewkowicz, Soto-Faraco et Sebastián-Gallés (2009), qui testaient des bébés ainsi que des adultes monolingues hispanophones sur leur capacités à discriminer le contraste /b/-/v/. Même si leur but était principalement d'observer un affinage perceptif multimodal, leurs résultats semblent intéressants à discuter au regard des nôtres. Dans leur étude, ils familiarisaient des bébés (ainsi que des adultes) anglophones et hispanophones avec deux vidéos côte-à-côte montrant une jeune femme articulant silencieusement /ba/ et /va/ afin de constituer une *baseline*. Par la suite, les bébés entendaient soit /ba/ soit /va/, à la suite de quoi il leur était de nouveau présenté les deux vidéos silencieuses. Comme preuve de la discrimination, les auteurs s'attendaient à ce que les bébés regardent préférentiellement (comparé à la *baseline*) la vidéo qui montrait une articulation cohérente avec la syllabe qu'ils avaient précédemment entendue.

Le pattern qui se dessine est le même pour les deux populations de six mois avec un regard préférentiellement tourné vers l'articulation qui coïncide avec la syllabe présentée. Les participants adultes hispanophones obtenaient le même pattern que les enfants de 11 mois, montrant que l'affinage perceptif visuel se maintient à l'âge adulte pour les hispanophones.

Ces résultats nous renseignent sur une limite de notre étude. En effet, les participants adultes de l'expérience précédemment citée ne parviennent pas à faire correspondre un /v/ oral avec sa réalisation articulatoire. Cela peut être dû soit (1) au fait que les participants ne distinguent pas les deux consonnes à l'oral, soit (2) au fait qu'ils ne distinguent pas les visèmes respectifs de chacune des consonnes. Cependant, l'étude de Pons et al. (2009) ne permet pas, au vue des données disponibles, de trancher la question. En effet, elle ne nous renseigne pas sur les capacités des participants à discriminer /ba/ et /va/ sur la base des informations auditives mais seulement sur la capacité à faire corresponde une syllabe orale à la réalisation articulatoire qui correspond. Aucun contrôle quant aux capacités à discriminer visuellement et auditivement ce contraste n'a été réalisé. Si le processus de correspondance échoue dans leur étude, cela peut donc être dû aussi bien à un mauvais encodage auditif que visuel de la labiodentale voisée. Dans le premier cas, il est envisageable que la moitié des essais de familiarisation auditive ayant été faite avec /va/, les participants (qui sont censés être sujets à une surdité phonologique), auraient lors de la familiarisation encodé /ba/ ou /fa/ (qui est régulièrement confondue avec /v/ par les hispanophones, Tomé, 1997) suite à une assimilation phonologique au lieu du /va/ qui était présenté. Cela resterait en contradiction avec nos résultats puisque, à population équivalente, nos hispanophones parviennent à discriminer /b/ et /v/ en modalité auditive. Nous pouvons donc supposer que c'est la correspondance visuelle qui pose problème aux participants dans l'étude de Pons et al. (2009).

Une contre-argumentation peut être avancée par le modèle de Hazan et al. (2006) qui postulent que les mouvements labiaux peuvent être utilisés s'ils sont partagés par les deux langues, ce qui est le cas de /v/ (puisque ce visème fait partie de la même classe que /f/ qui existe en espagnol). Cependant, ce modèle a été la plupart du temps testé avec des données obtenues avec des apprenants et non des individus monolingues. Comme dit précédemment, l'individu apprenant a déjà été familiarisé avec la L2, ce qui facilite peut-être l'utilisation des mouvements labiaux. De plus, le fait que nous n'ayons pas inclus de condition visuelle-seule dans notre étude nous empêche de conclure quant à l'utilisation des informations visuelles pour distinguer /v/ de /b/. En effet, il est envisageable que la réussite de notre groupe d'hispanophones lors de la présentation audiovisuelle soit uniquement liée à l'information

auditive. Les performances importantes observées lors de la présentation auditive ainsi que le manque d'une condition visuelle seule nous empêche de statuer sur la nature de l'information utilisée par les participants lors de la présentation audiovisuelle. Ainsi, il nous est impossible de savoir si les hispanophones sont capables de distinguer visuellement ces deux consonnes.

Enfin, la nature de la tâche utilisé par Pons et al. (2009), qui est sérielle (le signal auditif étant présenté avant l'apparition de l'articulation) est peut être une piste pour expliquer la discordance des résultats. Ce type de présentation discontinue ne fait pas appel au même processus qu'un présentation audiovisuelle simultanée puisque nous sommes ici plutôt dans le cas d'un *matching* inter-sensoriel et pas d'une intégration bimodale. En effet, même si les nouveau-nés et les adultes anglophones réussissent la tâche, l'échec des hispanophones reflète peut être simplement une difficulté à associer un événement sonore à un événement visuel présenté de manière séquentielle (une seconde d'intervalle) lorsque celui-ci implique un contraste non-natif. La réussite dans cette tâche est en effet conditionnée par des processus tel que la mémoire de travail, en plus des processus de discrimination. Ceux-ci sont peut-être mis en échec lors de la perception de phonèmes non natifs.

Nous avancerons pour notre part plusieurs explications pour comprendre l'absence de surdité phonologique observée chez les hispanophones. La première voudrait que ces deux phonèmes soient *assimilés* par les hispanophones à deux catégories phonologiques différentes, selon le « Perceptual Assimilation Model » de Best, McRoberts et Sithole (1988). En effet, il est envisageable que les participants catégorisent le /b/ français dans leur catégorie native de /b/, mais qu'ils perçoivent le /v/ comme étant plus similaire de leur /f/ que de leur /b/, ce qui faciliterait la distinction /v/(perçu comme /f/)-/b/. En effet, /v/ et /f/ ont une réalisation très similaire qui ne varie que sur le voisement (les deux consonnes étant des fricatives labiodentales). Une étude sur des participants hispanophones débutants en français irait dans ce sens. Tomé, en 1997, a fait écouter à des étudiants des triplets de mots français qui variaient sur la première consonne. Nous nous intéresserons ici seulement au triplet /b/-/v/-/f/. Leur tâche était de mettre une croix face aux mots dont ils jugeaient la prononciation similaire, et pouvaient ne rien cocher si tous les mots proposés étaient, selon eux, différents. Les interférences relevées dans les résultats sont plus importantes pour /v/-/f/ que pour /b/-/v/, c'est-à-dire que les participants avaient tendance à juger « fou » et « vous » comme étant le même mot, ce qui était moins le cas de « vous » et « bout ». Cela semble indiquer que les hispanophones assimileraient plus /v/ à /f/ qu'à /b/, ce qui concorde avec nos résultats. Il n'est pas exclu que le contraste étudié ne soit donc pas à même de faire émerger une surdité. Cela

nous amène à nous questionner sur la relativité du phénomène de surdité phonologique, qui serait non pas dépendante de l'inexistence d'un phonème dans le répertoire phonologique, mais dépendante du contraste dans lequel s'inscrit ce phonème. Le phonème /v/ pourrait en effet entraîner une surdité phonologique s'il est présenté dans le contraste /v/-/f/. D'autres études seront nécessaires afin de le déterminer.

Les résultats de Best & McRoberts (2003) nous fournissent également une autre piste. Ces auteurs suggéraient qu'une assimilation était observable seulement dans le cas d'utilisation d'articulateurs communs aux deux phonèmes et que la discriminabilité reste plus ou moins intacte dans le cas d'une place d'articulation différente. Selon eux, /b-v/ est un contraste *between-organ*, qui n'implique pas les mêmes articulateurs (ici une bilabiale et une labiodentale), ce qui permet de distinguer aisément ces deux phonèmes. Ces deux phonèmes peuvent être discriminés sur la seule base des informations auditives. Précisons tout de même que même si ce modèle expose des prédictions en accord avec nos résultats lorsque nous discutons du contraste /b/-/v/, il est cependant mis en échec si l'on considère la discrimination du contraste *between-organ* /f/-/θ/ qui n'est pas discriminé par les francophones lors de la présentation auditive, contrairement à ce que prédirait leur modèle.

Il se pourrait également que cette absence de surdité s'explique par à une exposition spécifique à une langue chez nos participants testés à Barcelone. Les participants sont en effet exposés à des réalisations de /v/ comme allophones de /f/ en position de coda ou lorsqu'il est suivi par une consonne comme c'est par exemple le cas dans le mot « afganistán » [avɣanistán] (Prieto, 2004).

Une explication d'ordre méthodologique est quant à elle peu envisageable. En effet, les stimuli utilisés dans les deux parties de l'étude ont été traités et enregistrés dans les mêmes circonstances. Ceci, associé au type de design expérimental ainsi qu'au nombre de stimuli utilisés, fait que les chances d'observer des résultats dus à un jugement purement acoustique et/ou visuel sont faibles. L'absence de surdité phonologique n'est donc pas le résultat d'un simple *matching* acoustique.

Une troisième cause peut expliquer l'absence de surdité mais également le fait que les temps de réaction des hispanophones sont plus élevés que ceux des francophones. Ce pattern de TRs est bien celui que nous attendions, cependant, il est important de discuter des causes. Cette observation pourrait nous faire croire que les hispanophones éprouvent des difficultés lors de la perception d'un contraste qui n'existe pas dans leur langue, ce qui ralenti leur TRs. Cependant, à la lumière de la première étude, nous pouvons tempérer ces propos. En effet les

résultats des hispanophones étaient déjà supérieurs à ceux des francophones lorsqu'il s'agissait de percevoir un contraste natif révélant une certaine « lenteur » de ce groupe. L'analyse des questionnaires a permis de mettre en évidence le fait que si les participants hispanophones devaient être monolingues (point important du recrutement dans une région de l'Espagne où le catalan est communément utilisé), certains d'entre eux se sont attribués des notes assez élevées (7 à 10 sur une échelle de 10 échelons) dans la maîtrise de langue autre que le catalan et l'espagnol (i.e., allemand (1=8), euskara (1=10), galicien (3=10), anglais (3=8 ; 5=7)). Cette « souplesse » dans le recrutement était due à la difficulté de l'expérimentateur de Barcelone pour recruter des participants monolingues « pures » pour la raison précédemment citée. Il sera sans doute nécessaire de tester quelques participants supplémentaires afin de remplacer les plurilingues. Cependant, cet élément pourrait expliquer les résultats. En effet, dans une autre étude menée sur des bilingues de la même région, nous avons également obtenu des temps de réaction plus importants que ceux des monolingues. Il semblerait que le traitement phonologique exige un coût supplémentaire pour les participants « plurilingues » et/ou bilingues (cf. Etude 3).

3.5 CONCLUSION GENERALE

La première contribution de cette étude est de mettre en évidence une utilisation efficace des mouvements labiaux pour discriminer les deux phonèmes alors même que ni le phonème, ni le visème qui lui est associé est connu des francophones. Le résultat inverse est couramment obtenu dans la littérature, à ceci prêt que la population étudiée est généralement constituée d'apprenants alors que notre étude porte sur des individus naïfs. De plus, nous observons un pattern différencié sur un même phénomène en fonction du contraste phonologique étudié. En effet, alors que le contraste /f/-/θ/ ne peut être discriminé sur la base des informations auditives par des francophones, le contraste /b/-/v/ qui n'existe pas en espagnol est quant à lui parfaitement discriminé. Les participants hispanophones ne semblent pas sujets à une surdité phonologique pour le contraste /b/-/v/. Les résultats du groupe contrôle d'hispanophones sont surprenants d'une part à cause de cette absence de surdité phonologique mais également car leur temps de réponses sont plus importants que ceux des francophones, quelque soit le contraste testé. En effet, cette tendance s'observe même lorsque les hispanophones perçoivent un contraste natif. Ces résultats restent difficiles à expliquer. Une exposition importante à

deux langues (espagnol et catalan) dans la vie de tous les jours ainsi qu'une mauvaise sélection des participants (plurilingues et non monolingues pour beaucoup) restent les meilleurs raisons pour expliquer de tels résultats.

3.6 RESUME

But de l'étude : Déterminer l'impact de la présentation audiovisuelle de la surdité phonologique dans le cadre de la perception de phonèmes inconnus perçus par des individus monolingues.

Populations : 20 francophones testés à Grenoble, France et 20 hispanophones testés à Barcelone, Espagne.

Protocole : Les deux contrastes étudiés étaient /f/-/θ/, qui n'existe pas en français et /b/-/v/ qui n'existe pas en espagnol. Les consonnes étaient insérées dans des séquences monosyllabiques ou bisyllabiques. Le paradigme expérimental consistait en la présentation d'un « mot de référence » qui devait être mémorisé au début du bloc, puis être comparé aux « mots » présentés par la suite. Pour chacun de ces stimuli, le participant devait répondre si ce « mot » était le même ou différent du mot de référence.

Résultats : Concernant les francophones, une surdité phonologique a été observée lors de la présentation auditive de séquences contenant le phonème non natif /θ/. La présentation des mouvements labiaux permet une amélioration des performances, qui deviennent similaires à celles des individus natifs. Un pattern différent est obtenu pour les hispanophones puisque ceux-ci ne semblent pas avoir de difficulté à distinguer /b/ et /v/, malgré le fait que ce dernier ne fasse pas parti de leur répertoire phonologique. Les hispanophones présentent également des temps de réactions supérieurs à ceux des francophones quels que soient la condition et le contraste.

Conclusion : Nous avons mis en évidence que les informations labiales peuvent être utilisées pour améliorer la perception des phonèmes 1) même sans connaissance préalable de la langue et 2) même si le visème en question est inconnu des participants.

CHAPITRE 4

ETUDE 2

MODULATION DES ACTIVATIONS NEURONALES LIEES A LA PERCEPTION AUDIOVISUELLE DE PHONEMES NON NATIFS

3.7 INTRODUCTION

Il est aujourd'hui admis que la vision améliore la compréhension de la parole. De nombreuses étude comportementales ont permis de mettre à jour une intégration des informations visuelles et auditives (McGurk & MacDonald, 1976), celle-ci permettant une augmentation des performances ainsi qu'une accélération les traitements, que les conditions d'écoute soient bruitées (Christian Benoît et al., 1994 ; Binnie et al., 1974 ; MacLeod & Summerfield, 1987 ; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007 ; Sumby & Pollack, 1954) ou non (Cerrato, Leoni, & Falcone, 1998 ; Grant & Seitz, 2000). Cependant, les mécanismes neuronaux impliqués lors de la perception audiovisuelle, et notamment ceux de l'intégration ne sont pas encore bien compris. La plupart des études menées sur la perception audiovisuelle étaient jusqu'à récemment menées en IRMf. Le but était principalement de déterminer les structures impliquées lors de la perception audiovisuelle, notamment lors l'intégration des deux signaux (Möttönen, Krause, Tiippana, & Sams, 2002 ; Pasley et al., 2012 ; Sams, Möttönen, & Sihvonen, 2005).

L'investigation réalisée en EEG a permis d'aborder les questions temporelles relatives à l'intégration, et notamment celles de l'impact des informations visuelles sur les traitements réalisés par le cortex auditif. Alors que les premières études réalisées montraient en général des intégrations audiovisuelles précoces émergeant entre 150 et 250 ms (Colin et al., 2002 ; Möttönen et al., 2002 ; Sams et al., 1991, 2005), d'autres études ont montré une intégration encore plus précoce qui affecte les traitements auditifs à partir de 100 ms (Besle, Fort, Delpuech, & Giard, 2004 ; Colin, Radeau, Soquet, Dachy, & Deltenre, 2002 ; van Wassenhove, Grant, & Poeppel, 2007 ; Wassenhove, Grant, & Poeppel, 2005). Contrairement aux études en IRMf qui attestent d'une activation plus importante des régions impliquées dans l'intégration lors d'une présentation bimodale comparée à la présentation unimodale (Calvert et al., 1999 ; Calvert, Campbell, & Brammer, 2000 ; Okada, Venezia, Matchin, Saberi, & Hickok, 2013), les études en EEG/MEG obtiennent généralement le pattern inverse, à savoir une diminution de l'activité neuronale lors de la présentation bimodale. Ces différences sont dues essentiellement au fait que les études en IRMf se focalisent généralement sur les sites multisensoriels car la faible résolution temporelle de cette technique ne permet pas d'observer des activations très précoces, alors que les études en EEG ou MEG se basent sur les activations précoces des aires sensorielles unimodales (Raij, Uutela, & Hari, 2000 ; Bushara et al., 2003).

Des études en EEG fournissent ainsi des pistes pour comprendre la modulation des processus auditifs par les informations visuelles des gestes de parole. Elles ont observé l'évolution temporelle des processus perceptifs au travers des potentiels évoqués (ERP : *Event Related Potentials*). Les potentiels évoqués sont des variations de l'activité électrique cérébrale en réponse à un stimulus donné. Ils sont extraits de l'électroencéphalogramme en moyennant les signaux produits par cette stimulation. Les potentiels évoqués sont généralement identifiés par leur polarité (positive ou négative), la latence (temps de déclenchement de la réponse par rapport au début de la stimulation, mesuré en millisecondes), l'amplitude (mesuré en μV) ainsi que la distribution topographique des modulations électrophysiologiques à travers le scalp (Coles & Rugg, 1995). L'amplitude reflète l'intensité des réponses neuronales, alors que la latence fait référence au moment où l'on observe celles-ci. Les pics précoces fournissent donc des informations sur les traitements qui sont réalisés au tout début du processus alors que les pics plus tardifs rendent compte des mécanismes opérant à des étapes subséquentes du traitement perceptif. Comme nous l'avons vu dans la partie théorique de ce manuscrit, quelques travaux sur la perception de la parole audiovisuelle ont étudié les potentiels évoqués auditifs précoces, c'est-à-dire, des processus auditifs ayant lieu dans les toutes premières millisecondes du traitement de l'information visuelle et audiovisuelle. Les potentiels auditifs précoces consistent en une séquence de pics connus comme le complexe N1/P2. Ce complexe est produit par une stimulation auditive qui se matérialise par un pic négatif autour de 100 à 150 ms (N1) et un pic positif dans les 200 à 250 ms (P2) après le début de la stimulation. Le rôle fonctionnel de N1 est en rapport avec les processus permettant la détection et l'encodage des propriétés auditives de la stimulation. Sa source neuronale a été localisée dans le cortex auditif primaire (Eggermont & Ponton, 2002 ; Näätänen & Picton, 1987). Le rôle de P2 est moins clair. Il est souvent décrit comme une période de positivité après l'occurrence de N1 et faisant partie du pattern d'activations. Il est possible que l'amplitude de P2 soit modulée par des facteurs contextuels, comme par exemple l'attention.

Quelques études en EEG/MEG ont permis de montrer des effets de suppression induits sur le cortex auditif par les informations visuelles lors de la présentation des mots (Shahin, Kerlin, Bhat & Miller, 2012), de syllabes et de voyelles (Arnal, Morillon, Kell, & Giraud, 2009 ; Besle et al., 2004; Klucharev, Möttönen, & Sams, 2003 ; Stekelenburg & Vroomen, 2007 ; Wassenhove et al., 2005) ainsi que de lettres de l'alphabet (Raij et al., 2000). Besle et al. (2004) ainsi que van Wassenhove et al. (2005, 2007) ont mis en évidence une modulation

des potentiels évoqués auditifs N1 pour les premiers et du complexe N1/P2 pour les seconds, liée à la modalité de présentation, dans une tâche de détection de cibles syllabiques. Par exemple, Besle et al. (2004) ont observé une interaction des modalités autour de 120-190 ms qui s'exprime par une diminution de la N1 auditive par rapport à la somme des activations générées par la présentation auditive et visuelle (Figure 20).

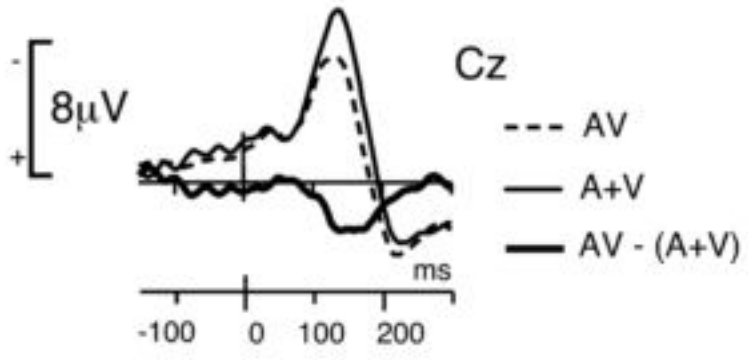

Figure 18. Réponse bimodale vs la somme des réponses unimodales de -150 à +300ms pour l'électrode Cz. (Extrait de Besle, Fort, Delpuech, & Giard, 2004)

Ils obtiennent également une accélération des réponses comportementales lors de la présentation audiovisuelle, mais aucun avantage temporel n'est observé sur les signaux EEG.

D'autres études ERP sur la perception de la parole audiovisuelle indiquent en revanche que la détection visuelle des mouvements de préparation articulatoire, ajoutée à celle de la détection auditive, accélère également le déroulement temporel des traitements acoustiques (Stekelenburg & Vroomen, 2007). En particulier, l'étude de van Wassenhove et al. (2005) a mis en évidence que les informations visuelles fournies par les gestes labiaux « préparateurs » en présentation audiovisuelle - par exemple, les mouvements labiaux qui précèdent la fermeture labiale afin de produire le phonème /p/ dans la syllabe /pa/ - diminue les latences des composantes évoquées N1 et P2 au niveau des aires auditives (Figure 21).

Dans une présentation visuelle seule, ils ont observé des potentiels à environ 400 ms avant le début acoustique du stimulus qui n'ont pas donné lieu à un potentiel évoqué auditif ; ils ont observé des potentiels évoqués visuels typiques dans les zones temporo-occipitales. Il apparaît également que la magnitude du décalage dans la latence du complexe N1/P2 dépend de la saillance visuelle du geste de production. Par exemple, /p/ (comme dans /pa/), qui est produit avec un geste labial, est plus saillant visuellement que /k/ (comme dans /ka/), qui est

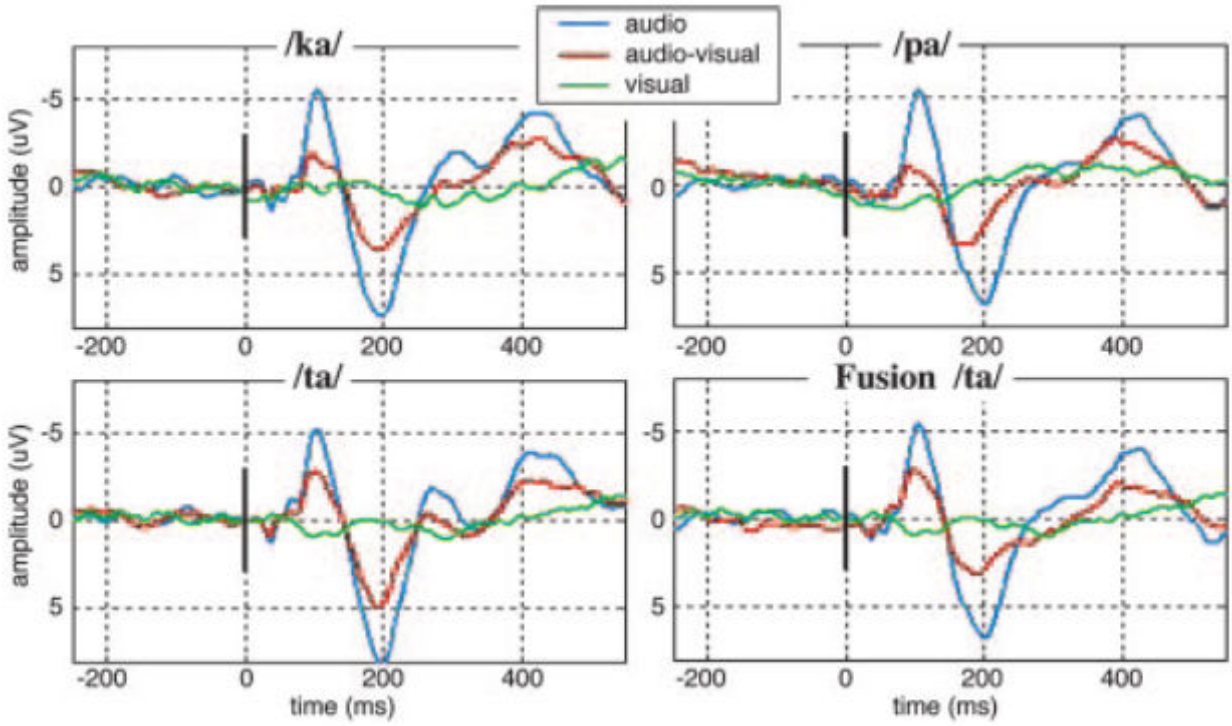

Figure 19. Potentiels évoqués moyens obtenus pour les quatre types de stimuli (i.e., /p-t-k-fusion/) en fonction de la modalité de présentation (A, AV et V) obtenus sur l'électrode centro-pariétale. La ligne verticale indique le début du signal auditif. (Extrait de van Wassenhove et al., 2005)

produit grâce à une occlusion de la partie postérieure de la cavité buccale. Les résultats des ERP montrent que /p/ produit des décalages de latence en N1 et P2 plus importants que /k/.

La suppression de la N1 auditive ou désactivation corticale serait liée à une facilitation du traitement des indices acoustiques, qui serait par conséquent plus rapide (Besle et al., 2004). Elle serait due au fait que les informations oro-faciales, qui précèdent les informations auditives (à cause de la coarticulation anticipatoire), réduirait l'incertitude du signal et diminuerait les demandes computationnelles des aires du cortex auditif (Besle et al., 2004; Wassenhove et al., 2005). Est-ce que ces modulations de l'activité neuronale du cortex auditif par l'information visuelle peuvent expliquer l'amélioration des performances lorsque nous devons établir une conversation dans une langue étrangère ? Nous avons vu dans les chapitres précédents que l'information visuelle sur les gestes de production du locuteur peut même nous permettre de surmonter la surdité phonologique dont nous sommes victimes en situation auditive seule, mais que se passe-t-il au niveau neuronal qui pourrait expliquer cette amélioration ? Des ondes précoces telles que N1 et P2 sont peut être déjà sujettes à des modulations lors de la présentation de mouvements préparatoires inconnus et moduleront peut-être les traitements phonologiques subséquents, plus tardifs.

3.7.1 OBJECTIFS

Le traitement de phonèmes français et des phonèmes d'autres langues seront évalués selon les différentes configurations expérimentales (audio seul, visuel seul et audiovisuel) à l'aide de mesures comportementales (pourcentage de réponses correctes) et de l'activité électro-encéphalographique. Les processus neuronaux mis en jeu seront mesurables en termes de modulation d'amplitude et de latence des potentiels évoqués. Le but principal de l'étude est d'observer les modulations des composantes auditives précoces en fonction de l'existence ou non d'un phonème dans notre langue. En effet, les études précédemment citées se sont uniquement intéressées à la perception des phonèmes natifs. Cependant, il n'est pas acquis que la perception visuelle d'un mouvement inconnu entraîne forcément une modulation des composantes auditives N1/P2. En effet, alors qu'une suppression est observée aussi bien pour des événements langagiers (dont les conséquences sont connues) (Arnal et al., 2009 ; Besle et al., 2004 ; Colin et al., 2002 ; van Wassenhove et al., 2007 ; Wassenhove et al., 2005) que non langagiers (Stekelenburg & Vroomen, 2007), nous ne pouvons pas assumer que cela soit le cas lors de la perception de mouvements articulatoires inconnus. En effet, si la diminution de l'amplitude de N1/P2 est liée à des prédictions permettant de diminuer les demandes computationnelles du cortex auditif, est-ce que des mouvements aux conséquences acoustiques inconnues permettront une telle réduction ? La N1 devrait être sujette à une réduction d'amplitude et/ou de latence puisque celle-ci est modulée à partir du moment où des mouvements anticipatoires sont présents, même si les éléments présentés ne sont pas congruents (Stekelenburg & Vroomen, 2007). Cela devrait donc s'appliquer aux mouvements articulatoires inconnus. En revanche, la P2, qui est influencée par la congruence, sera peut être modulée lors de la perception audiovisuelle d'un phonème non natif.

Nous avons choisi d'utiliser la consonne espagnole interdentale fricative non voisée /θ/ car elle n'existe pas en français et son articulation présente des indices visuels très saillants qui seront les plus à même d'être exploités par l'auditeur (Etude 1). Elle sera contrastée avec la consonne labiodentale fricative non voisée /f/ qui existe aussi bien en français qu'en espagnol. Nous faisons l'hypothèse que le traitement visuel des gestes labiaux lors de la perception des phonèmes non natifs va moduler les composantes N1 et P2 au niveau des aires auditives. Nous allons tester une population de francophones natifs. Ces résultats seront comparés à ceux d'un groupe de contrôle d'hispanophones natifs. Les questions auxquelles nous souhaitons répondre dans cette étude sont :

- Est-ce que la présentation audiovisuelle va réduire l'amplitude et la latence des composantes auditives N1/P2 pour le phonème /f/ chez les francophones et les hispanophones ?
- Est-ce que la présentation audiovisuelle va réduire l'amplitude et la latence des composantes auditives N1/P2 pour le phonème /θ/ chez les francophones – pour qui le phonème est non natif - et les hispanophones ?

3.8 MATERIEL ET METHODE

3.8.1 PARTICIPANTS

Le groupe expérimental était composé de 19 francophones monolingues (15 femmes; *M* = 23;1 ans, SD = 3.56) et de 13 hispanophones monolingues originaires d'Espagne (9 femmes; *M* = 25 ans, SD = 3.56). Tous les participants étaient droitiers, avaient une vision normale ou corrigée à la normale. Ils ne présentaient pas de problèmes d'audition ni de trouble langagier ou moteur. Un consentement écrit a été signé par tous les participants. L'étude a été approuvée par le Comité de Protections des Personnes (CPP) Sud-Est 5, n° ID RCB 2013-A00123-42. Les participants francophones n'avaient pas de connaissance antérieure de l'espagnol.

3.8.2 MATERIEL

Dans cette étude, nous avons étudié le contraste consonantique espagnol /θ/-/f/. Nous présenterons à des participants francophones et hispanophones natifs des syllabes espagnoles contenant un phonème qui existe en français (e.g., /fa/ : natif) et un phonème inexistant en français (e.g., /θa/ : non natif) afin de montrer qu'ils les confondent en présentation auditive seule (cf. notion de surdité phonologique) et qu'ils arrivent à les distinguer en présentation audiovisuelle. La Figure 22 présente des images de l'articulation des phonèmes utilisés.

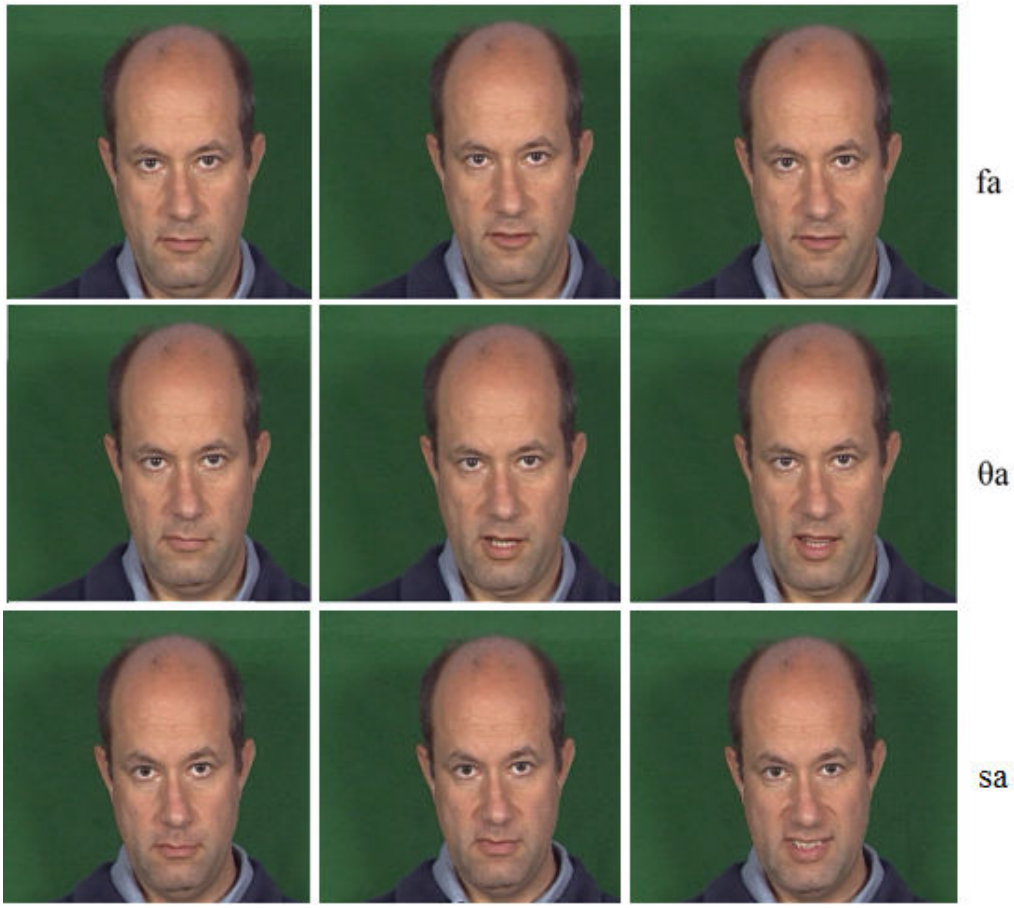

Figure 20. Représentation des trois premières trames contenant les premiers mouvements articulatoires de la prononciation des stimuli /fa/, /θa/ et /sa/.

Les stimuli monosyllabiques de l'étude 1 ont été utilisés. Les exemplaires de /fa/, /θa/ et /sa/ étaient prononcés par un locuteur hispanophone natif (qui n'a pas participé à l'expérience) dans une chambre sourde. Les séquences ont été enregistrées avec un microphone AKG 1000S et une caméra vidéo numérique de haute qualité placée devant le locuteur. Le flux audio des stimuli a été échantillonné à 44.1 kHz et la vidéo était présentée au format 688x572 pixel/images à un taux de rafraichissement de 50Hz. La syllabe /fa/ était utilisée pour les stimuli natifs, c'est à dire qui existe dans le répertoire phonologique des participants

francophones. La syllabe /θa/ a été utilisée comme stimulus non natif car la fricative interdentale n'existe pas en français. La syllabe /sa/ existe en français et était utilisée comme stimulus natif contrôle.

Les stimuli utilisés ont été sélectionnés en fonction de la durée de leur friction (le maximum de paramètres doivent être communs pour une expérimentation en EEG surtout dans le cas d'hypothèse sur la latence des ondes générées). Nous avons donc sélectionné deux exemplaires du stimulus composés du phonème natif présentés (i.e., /fa/) (six autres exemplaires étaient utilisés comme « référence ») ; deux stimuli « non-natifs » qui contiennent un phonème qui est confondu avec le phonème consonantique de la référence (e.g., /θa/) et deux séquences que nous avons appelés « natif contrôle » qui contiennent un phonème consonantique qui n'est pas confondu avec le phonème consonantique de la référence et/ou un phonème vocalique différent (e.g., /sa/ : natif contrôle).

Tous les stimuli commençaient par un fade-in de 200ms qui était suivit par un *jitter* qui variait entre 800 et 1200 ms. Le fade-in et le *jitter* ont été créés en répliquant la première image de la vidéo en utilisant Adobe Premiere. La séquence d'intérêt (du début des mouvements articulatoires jusqu'à la fin du signal audio) faisait 520ms (Figure 23). Elle consistait en une partie silencieuse d'environ 200 ms (+/- 20ms) durant laquelle seuls les mouvements articulatoires étaient visibles.

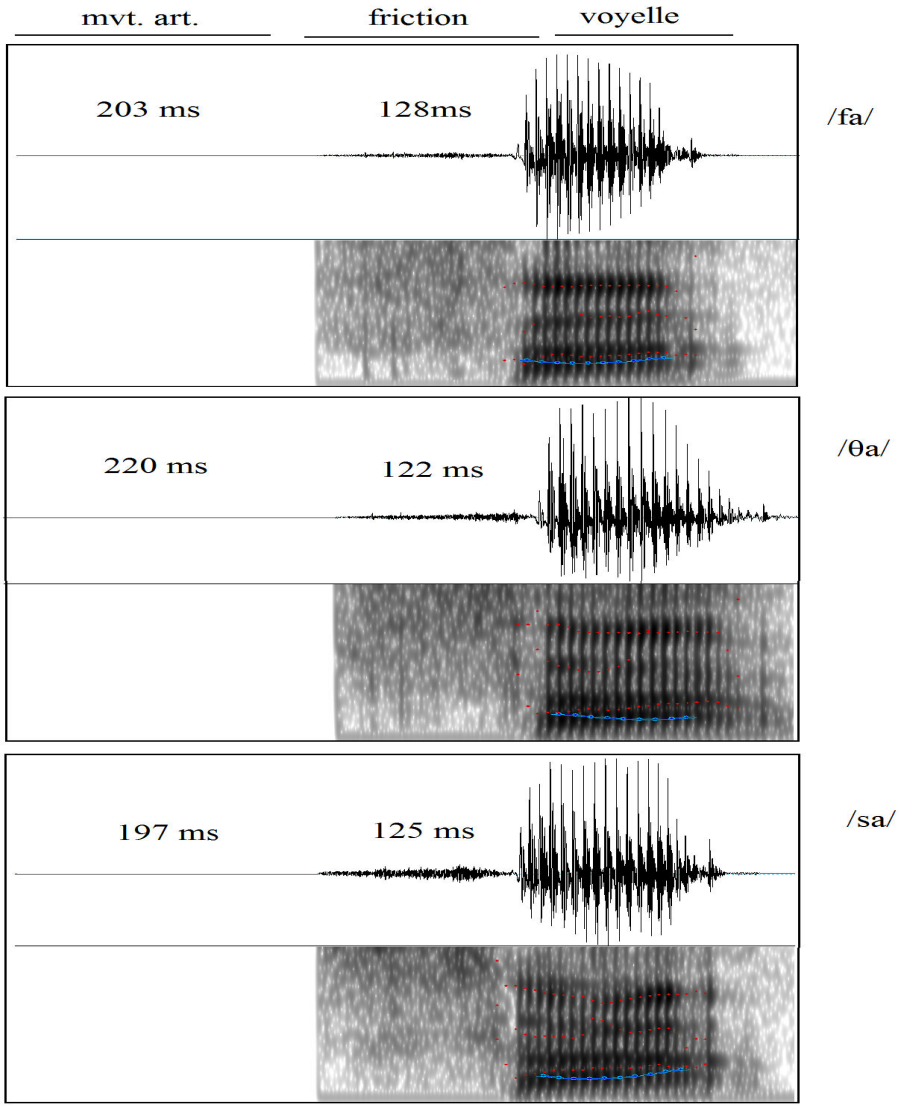

Figure 21. Illustration du signal acoustique et spectrogramme (un des deux exemplaires utilisé dans l'étude) pour les séquences /fa/, /θa/ et /sa/. Pour chacune d'elles, la durée du début de la séquence (premiers mouvements articulatoires) jusqu'a la friction, ainsi que la durée de la friction sont indiquées.

Notons que les mouvements articulatoires commençaient au même moment quelle que soit la séquence, soit une trame après le début de la séquence en mouvement. L'onset audio (OA) coïncidait avec le début de la friction qui durait 125 ms en moyenne (+/- 4ms). Elle était suivie par la voyelle /a/. A la fin du stimulus, un délai de 500ms a été ajouté avant l'écran de réponse afin de limiter les « contaminations » induites par la réponse motrice. Les stimuli auditifs étaient réalisés en dupliquant la première image de la vidéo jusqu'à la fin du signal audio. Les stimuli visuels étaient les mêmes que les stimuli audiovisuels, mais la bande son était supprimée.

3.8.3 PROCEDURE

La tâche comprenait 18 stimuli (3 syllabes * 2 exemplaires * 3 jitters) répétés 14 fois pour un total de 252 essais. Après avoir expliqué la tâche aux participants, la session expérimentale était divisée en 3 blocs, un pour chaque modalité de présentation (auditive, audiovisuelle et visuelle). Les blocs de présentation auditive et audiovisuelle étaient contrebalancés entre les sujets et la modalité visuelle était toujours la dernière à être réalisée. La bande son était diffusée par deux enceintes M Audio Studiophile AV 30 placées de chaque côté d'un écran Dell de 22 pouces à un taux de rafraichissement de 60Hz. Au début d'un bloc, le participant était exposé à six différents exemplaires du « mot » de référence /fa/, qui était toujours présenté en audiovisuel. Les participants devaient écouter et regarder attentivement la vidéo dans le but de mémoriser le « mot ». Le terme « mot » était utilisé tout au long des consignes afin de maximiser l'encodage phonologique de la référence. Après avoir été exposé à la référence, des points d'exclamation apparaissaient durant 2000 ms pour indiquer au participant qu'il/elle allait être exposé/e à la liste complète de stimuli. Dans cette liste /fa/, /θa/ et /sa/ apparaissaient de manière aléatoire et dans la même proportion (33,33%). Les participants avaient 2000 ms pour décider si le "mot" qu'ils venaient de percevoir était le même ou un "mot" différent de celui qu'ils avaient mémorisé en référence, en pressant la flèche droite ou gauche d'un clavier situé devant eux. Les boutons de réponse étaient contrebalancés entre les sujets. Un bloc durait environ 17 min. Avant de commencer l'expérience, les participants faisaient un entraînement en présentation audiovisuelle de six essais pour s'assurer qu'ils comprennent bien la tâche et un autre avec les mêmes stimuli afin de leur présenter les consignes spécifiques à l'EEG. Des stimuli différents de ceux utilisés dans l'expérience étaient présentés durant les entraînements.

Avant chaque bloc (A, AV ou V) le participant avait un signal lui indiquant qu'il devait se détendre et ne penser à rien pendant 30 secondes, afin que l'on puisse enregistrer l'activité cérébrale de repos.

Le déroulement de chaque essai est présenté sur la Figure 24.

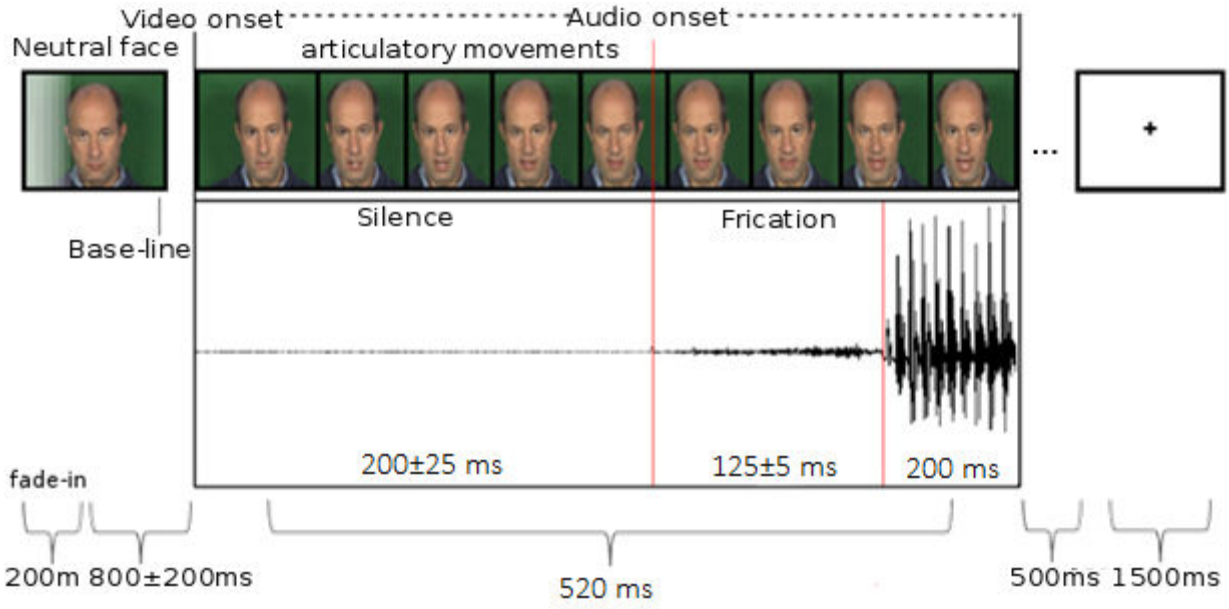

Figure 22. Design expérimental d'un essai.

L'essai débutait par une croix de fixation de 400 ms. Un visage neutre apparaissait ensuite progressivement sur 200 ms et était maintenu sur une durée de 800 +/- 200 ms. La syllabe était ensuite présentée sur 600 ms et le participant avait 1500 ms pour répondre. La durée maximale d'un essai était donc de 3700 ms. Il y avait 84 essais pour chaque type de stimuli (natif, non natif, natif contrôle) par modalité de présentation (ceux-ci ne comprennent pas les 6 stimuli « référence » qui seront présentés au début de chaque liste). Cela implique 84 x 3 (types stimuli), soit une durée maximale de 18,5 min par bloc (60 min au total pour les trois modalités). Nous avons mesuré le taux de réponses correctes à l'aide du logiciel de Présentation.

3.8.4 ACQUISITION EEG

L'activité électroencéphalographique (EEG) du participant a été enregistrée en continu pendant toute la durée de la séquence expérimentale à l'aide d'un casque EEG BrainCapTM actiCAPTM de 64 96 électrodes actives (produit de la société Brain Products GmbH, agréé CE recherche et médical) positionnées selon le système standard 10/20. Le signal EEG, filtré à travers un passe-haut (0.01 Hz), a été échantillonné à une fréquence de 512Hz. Les électro-oculogrammes mesuraient les mouvements horizontaux (EOGH) et verticaux (EOGV) des

yeux en utilisant des électrodes externes placées aux coins externes des yeux, au-dessus et en dessous de l'œil gauche. Avant l'expérience, l'impédance de toutes les électrodes était ajustée de façon à être inférieure à 5 kOhms.

3.9 ANALYSES DE DONNEES

3.9.1 DONNEES COMPORTEMENTALES

Nous avons calculé le pourcentage de réponses correctes pour chaque participant, stimuli et modalité sur la base du taux de réponse.

3.9.2 ANALYSE EEG

Les données EEG ont été traitées en utilisant la Toolbox Fieldtrip pour l'analyse des signaux EEG/MEGEEG (FC Donders Centre for Cognitive Neuroimaging, Nijmegen, The Netherlands;) via Matlab (Mathworks, Natick, MA, USA). Les signaux ont été segmentés en utilisant une fenêtre de 1000 ms centrée sur l'*onset* audio de chaque essai. Le signal a ensuite été re-référencé par la moyenne de toutes les électrodes, après avoir visuellement retiré les électrodes pour lesquelles de forts artefacts électriques étaient présents sur tout l'enregistrement (i.e., moins de 9% de toutes les électrodes de tous les participants). Les essais déviants ont été supprimés par le biais d'une procédure en deux étapes qui incluait un rejet visuel basé sur la variance et les valeurs maximales absolues de chaque essai, ainsi qu'une procédure semi-automatique basée sur les canaux EOG pour retirer les essais comportant des artefacts oculaires verticaux ou horizontaux (voir REF pour une explication détaillée des méthodes). Une moyenne de 9.28% +/- 3% des essais ont été supprimés pour la condition audiovisuelle, 13.96% +/- 15% pour la condition auditive et 7.77% +/- 3% pour la condition visuelle. Les canaux manquants ont été reconstruits en utilisant une interpolation par *spline* sphérique. Les potentiels évoqués ont ainsi été obtenus pour chaque type de stimuli et modalité en moyennant les essais correspondants, avec une correction de *baseline* de -200 ms à 0 ms avant le début de la vidéo, et filtré à 2-30Hz. La grande moyenne et les topographies au niveau du groupe ont été calculées sur tous les sujets.

Comme les potentiels évoqués N1/P2 ont une réponse maximale au niveau des électrodes centrales (Näätänen & Picton, 1987; Scherg & von Cramon, 1986), et en accord avec de précédentes études sur la perception audiovisuelle de la parole et les potentiels évoqués

auditifs (Baart, Vroomen, Shaw, & Bortfeld, 2014 ; Baart & Vroomen, 2010 ; Pilling, 2009; Stekelenburg & Vroomen, 2007 ; Treille et al., 2014 ; van Wassenhove et al., 2005) les analyses statistiques se sont concentrés sur les trois électrodes centrales gauche, centrale et droite (C3, Cz, C4). Les analyses des différences groupes/conditions ont été réalisées à l'aide de tests non-paramétriques, basés sur l'estimation a posteriori de la distribution des probabilités (méthode Monte-Carlo). Ces tests étaient réalisés dans la fenêtre de -500 à 500 ms autour de l'*onset* auditif en procédant à 10 000 permutations pour chaque point temporel (échantillon). Ces comparaisons ne considèrent comme significatives que les périodes durant lesquelles la probabilité se maintenait à $p < .05$ durant au moins 15 échantillons temporels consécutifs, soit une période de 15*2ms= 30ms (Blair & Karniski, 1993; Thorpe, Fize, & Marlot, 1996).

3.10 RESULTATS

3.10.1 ANALYSE COMPORTEMENTALE

Une analyse de la variance (ANOVA) incluant le Groupe (francophones, hispanophones) en facteur intra-sujet et la Modalité de présentation (auditive, audiovisuelle, visuelle) et le Type de stimuli (natif, non natif, natif contrôle) comme facteurs inter-sujets a été réalisée sur le pourcentage de réponses correctes.

L'analyse a révélé un effet significatif de la modalité de présentation, $F_{(1,29)} = 9.42$, $p < .001$. Les résultats sont représentés dans la figure 25.

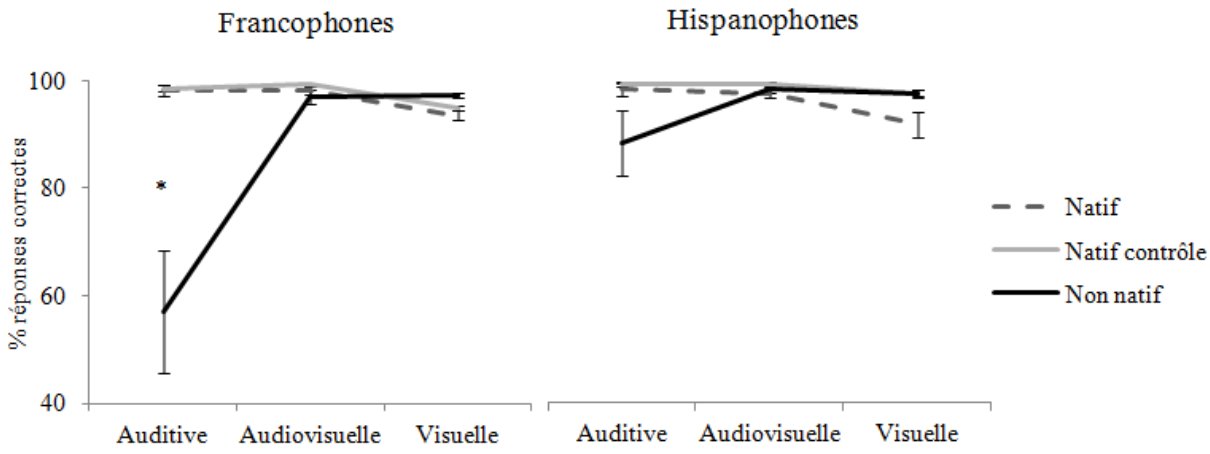

Figure 23. Pourcentage de réponses correctes pour le groupe de francophones (à gauche) et d'hispanophones (à droite) en fonction de la modalité de présentation (audiovisuelle, auditive, visuelle) et du Type de stimuli (natif, non natif, natif contrôle). * = p <. 05

Le score était plus important lors de la présentation audiovisuelle ($M_{audiovisuelle}$ = 98%, SD = 0.7), suivi de la modalité visuelle ($M_{visuelle}$ = 95%, SD =0,9), puis de la modalité auditive ($M_{auditive}$ = 89%, SD = 3,4). L'effet du Type de stimuli était également significatif, $F_{(1,29)}$= 13.85, p<.001. Les stimuli natifs étaient ceux pour lequel le score était le plus élevé (M natif contrôl = 98% ; SD = 0,4) suivi par les stimuli natifs contrôle (M_{natif} = 96% ; SD = 1,1) puis par les stimuli non natifs ($M_{non\ natif}$ = 89% ; SD = 3,4). L'effet principale du Groupe n'était pas significatif, $F_{(1,29)} = 2.33$, $p = .13$; les scores moyen des deux populations étaient équivalents.

L'interaction Modalité de présentation*Groupe était significative, $F_{(1,29)} = 4.78$, p<.05. Celle-ci est largement due à des performances plus faibles lors de la présentation auditive pour le groupe de francophones. La présentation bimodale permet une amélioration

significative du score des francophones de 13,5% (tous stimuli confondu), $F_{(1,29)} = 27.0$, $p < .001$. Le score des hispanophones n'augmente pas significativement ($F < 1$). L'interaction Modalité de présentation*Type de stimuli était significative, $F_{(1,29)} = 15.16$, $p < .001$; ainsi que l'interaction Type de stimuli*Groupe, $F_{(1,29)} = 6.19$, $p < .01$. Enfin, l'interaction triple était significative, $F_{(1,29)} = 4.74$, $p < .01$.

Pour le groupe de francophones, les comparaisons par paires révèlent que lors de la présentation auditive, il n'y a pas de différences entre les stimuli natif et natif contrôle, $F < 1$. En revanche et comme nous l'attendions, la différence était significative entre les stimuli natifs et non natifs, ($F_{(1,29)} = 26.51$, $p < .001$, et entre les stimuli non natifs et natifs contrôle, $F_{(1,29)} = 27.59$, $p < .001$. Lors de la présentation audiovisuelle, les scores entre les stimuli natifs et non natifs ($F_{(1,29)} = 1.59$, $p = .21$) et natifs et natifs contrôle ($F_{(1,29)} = 2.44$, $p = .12$) étaient équivalents. Les scores étaient plus importants pour les stimuli natifs contrôle que non natifs, $F_{(1,29)} = 5.54$, $p < .05$. Enfin lors de la présentation visuelle seule, aucune différence n'est observée entre les stimuli (natif vs. non natif, $F_{(1,29)} = 2.84$, $p = .10$; non natif vs. natif contrôle, $F < 1$; natif vs. natif contrôle, $F < 1$.

Durant la présentation des phonèmes natifs, une différence est observée entre les présentations audiovisuelle et visuelle, $F_{(1,29)} = 4.59$, $p < .05$. Les autres comparaisons ne sont pas significatives (auditive vs. audiovisuelle, $F < 1$; auditive vs. visuelle, $F_{(1,29)} = 3.41$, $p = .07$). Pour les phonèmes non natifs, alors qu'aucune différence n'est obtenue entre les scores en audiovisuelle et visuelle, $F < 1$, les comparaisons entre les modalités auditive et visuelle, ainsi qu'auditive et audiovisuelle sont significatives ($F_{(1,29)} = 27.34$, $p < .001$ et $F_{(1,29)} = 26.61$, $p < .001$ respectivement). Le gain obtenu lors de la présentation audiovisuelle du phonème /θ/ par rapport à la présentation auditive atteint 39,9%. Cela tend à annuler les différences entre les deux groupes lors de la présentation audiovisuelle du phonème non natif, $F < 1$. Enfin, aucune des trois modalités n'entraînent des différences dans les scores pour les phonèmes natifs contrôles (auditive vs. audiovisuelle, $F_{(1,29)} = 1,91$, $p = .17$; audiovisuelle vs. visuelle, $F_{(1,29)} = 1.82$, $p = .18$; auditive vs visuelle, $F_{(1,29)} = 1.27$, $p = .26$).

Lors de la présentation auditive, aucun des scores n'est modulé par le Type de stimuli (natif vs. non-natif, $F_{(1,29)} = 1.01$, $p = .32$; natif vs. natif contrôle, $F < 1$; non natif vs. natif contrôle, $F_{(1,29)} = 1.18$, $p = .28$). Le même pattern est obtenu lors de la présentation audiovisuelle (natif vs. non-natif, $F < 1$; natif vs. natif contrôle, $F_{(1,29)} = 3.42$, $p = .07$; non natif vs. natif contrôle, $F < 1$. Enfin, lors de la présentation visuelle, la différence entre les phonèmes natif et natif contrôle est la seule à être significative, $F_{(1,29)} = 9.80$, $p < .01$ (natif vs.

non natif, $F_{(1,29)} = 4.07$, $p = .05$; natif contrôle vs. non natif, $F < 1$).

Lorsque les hispanophones perçoivent le phonème natif /f/ en modalité visuelle, les scores sont moins élevés que ceux obtenus lors de la présentation auditive, $F_{(1,18)} = 4.55$, $p < .01$ et audiovisuelle, $F_{(1,18)} = 4.47$, $p < .01$. Les scores pour ce phonème sont identiques lors de la présentation audiovisuelle et auditive, $F_{(1,18)} = 1.13$, $p = 0{,}29$. Lors de la présentation du phonème non natif (qui est en fait natif pour les hispanophones), les scores ne diffèrent pas en fonction de la modalité de présentation, les scores étant équivalents entre les modalités audiovisuelle et auditive, $F_{(1,18)} = 1.05$, $p = .31$, audiovisuelle et visuelle, $F < 1$ et enfin, auditive et visuelle, $F < 1$. Le même pattern est obtenu pour les stimuli natif contrôle (auditive vs. audiovisuelle, $F < 1$; audiovisuelle vs. visuelle, $F < 1$ et auditive vs visuelle, $F < 1$).

Enfin les comparaisons entre les deux groupes ne révèlent qu'une seule différence. Celle-ci est obtenue, comme prédit par nos hypothèses, lors de la présentation auditive des stimuli non natifs, $F_{(1,18)} = 5.81$, $p < .05$. Les scores des deux groupes sont équivalents pour les autres comparaisons (auditive : natif, $F < 1$; natif contrôle, $F < 1$; audiovisuelle : natif, $F < 1$; non natif, $F < 1$; natif contrôle, $F < 1$; visuelle : natif, $F < 1$; non natif, $F < 1$; natif contrôle, $F < 1$.

3.10.2 ANALYSES EEG

3.10.2.1 IMPACT DE LA PRESENTATION AUDIOVISUELLE SUR N1 ET P2

3.10.2.1.1 FRANCOPHONES

Nous présenterons dans un premiers temps les topographies obtenues lors des présentations auditive et audiovisuelle pour les phonèmes natif, natif contrôle et non natif (Figure 26).

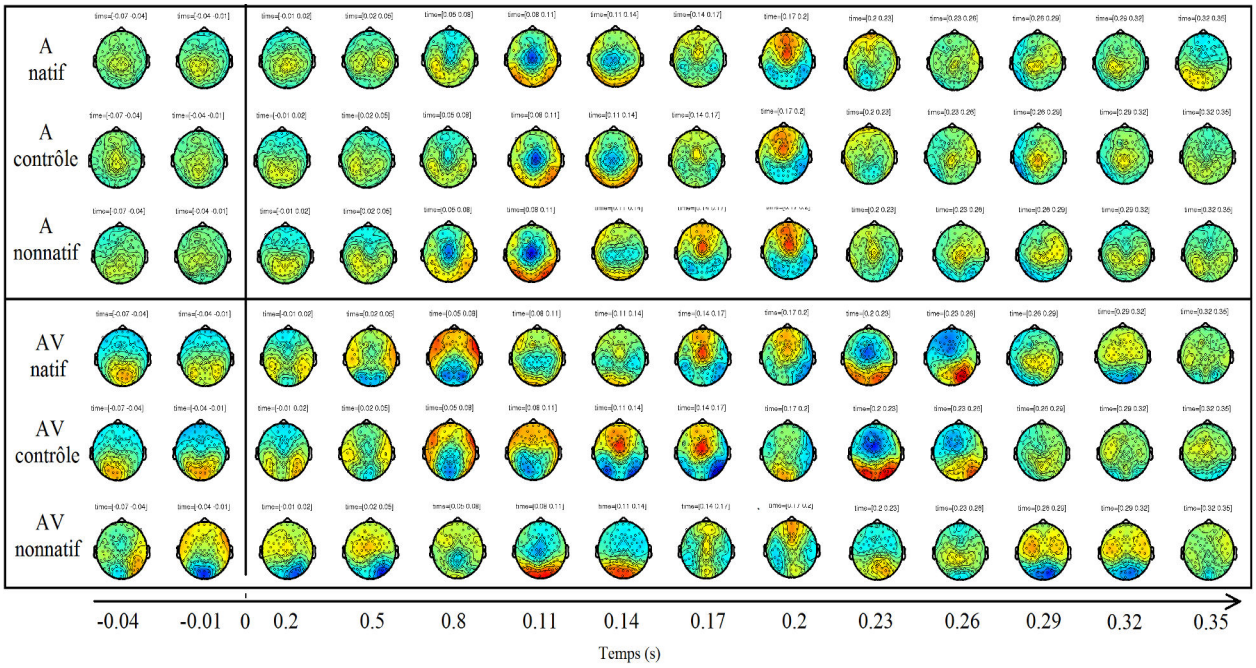

Figure 24. Topographies obtenues lors de la présentation auditive et audiovisuelle des phonèmes natif, natif contrôle et non natif pour le groupe de francophones.

Nous pouvons dans un premier temps observer, de manière descriptive, les patterns d'activation centraux classiquement évoqués lors de la présentation de stimuli auditifs, témoins entre autre de l'activation des aires auditives. Nous retrouvons également les patterns occipitaux et centraux (surtout pour le phonème natif) évoqués lors de la présentation audiovisuelle, témoins quant à eux de l'activation conjointe des aires auditives et visuelles. Les activations centrales sont plus précoces lors de la présentation auditive du phonème non natif par rapport à celles du phonème natif ou natif contrôle (environ à 80 et 110 ms respectivement). Les activations pour les deux stimuli natifs sont d'ailleurs similaires lors de la présentation auditive. Une activation centrale est également observée plus précocement lors de la présentation auditive du phonème non natif (170 ms) par rapport au stimulus natif (200 ms). Celle-ci semble plus étendue temporellement dans le cas des phonèmes non natifs, les deux activations s'atténuant autour de 230ms. Cependant, les patterns sont assez semblables. En revanche, de nombreuses différences sont observables lors de la présentation audiovisuelle des deux phonèmes. L'activité occipitale est observable de manière bien plus précoce lors de la présentation du phonème non natif (de -10 à 50ms) qui donne par la suite lieu à un potentiel central, peu intense, dont la polarité change de 20 à 110 ms après le début du signal acoustique. Ces activations centrales sont bien moins marquées lors de la perception audiovisuelle de phonèmes non natifs (pour lesquelles les activations occipitales sont prédominantes). Les activations fronto-temporale positive observée à 80 ms, ainsi que

centrale positive à 140-170 ms (observées à plus ou moins 140-170 ms sur toutes les autres cartes pour cette dernière) sont absentes lors de la présentation du phonème non natif.

Nous attirons l'attention sur le fait qu'il est très difficile, voire faux, d'inférer sur les aires corticales précisément activées directement à partir des topologies de scalp. Pour ce faire, il serait nécessaire d'utiliser des algorithmes de reconstruction de sources permettant de modéliser le signal au niveau des sources corticales.

Afin de vérifier notre première hypothèse concernant la diminution d'amplitude de la N1/P2 lors de la présentation audiovisuelle, nous avons comparé les potentiels évoqués lors de la présentation auditive et audiovisuelle en fonction du Type de stimuli pour les deux populations. Les analyses réalisées sur le groupe de francophones montrent une diminution de l'amplitude de la N1 pour les stimuli natifs et non natifs (Figure 27 a et b respectivement).

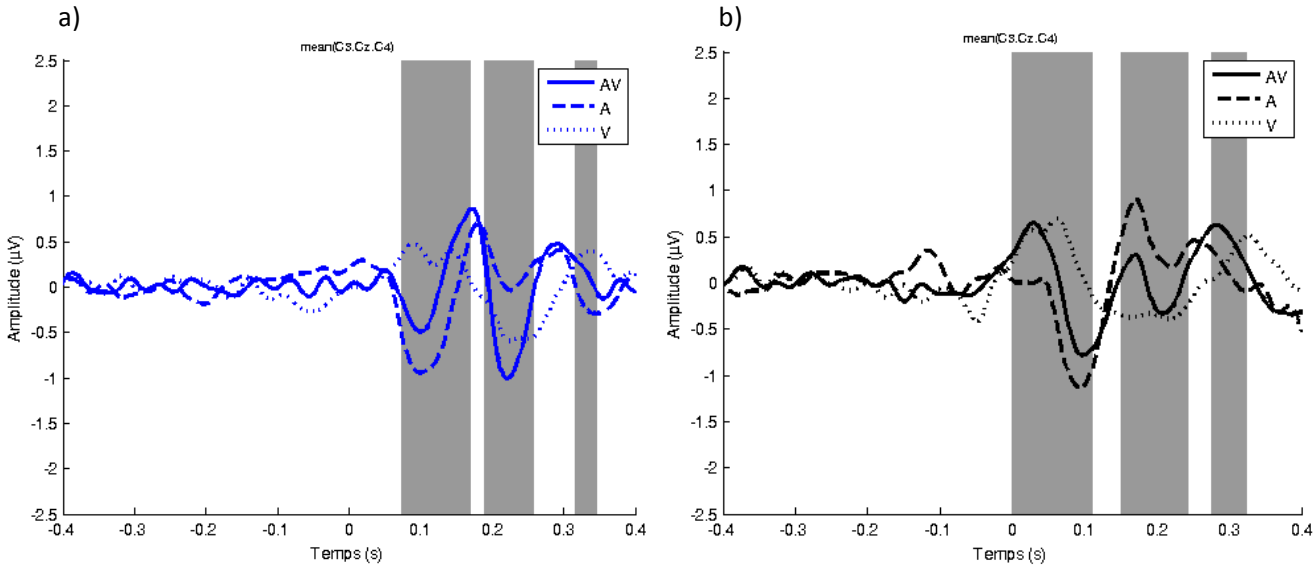

Figure 25. Potentiels évoqués lors de la présentation audiovisuelle, auditive, et visuelle pour les stimuli natifs (a) et non natifs (b). Les zones grisées représentent les différences entre la modalité audiovisuelle et auditive ($p < .05$)

Nous constatons également la présence de différences lors de la présentation des stimuli natifs contrôle (Figure 28 a). Cependant, les potentiels auditifs générés lors de la présentation auditive et audiovisuelle ne sont pas alignés, ce qui limite les inférences concernant l'amplitude de ces potentiels. Afin de s'assurer que les pics de N1 des deux modalités diffèrent en amplitude, nous avons procédé à un réalignement temporel des courbes, en décalant le signal généré lors de la présentation auditive de -30 ms. Cet ajustement effectué, nous avons reconduit les analyses (Figure 28 b).

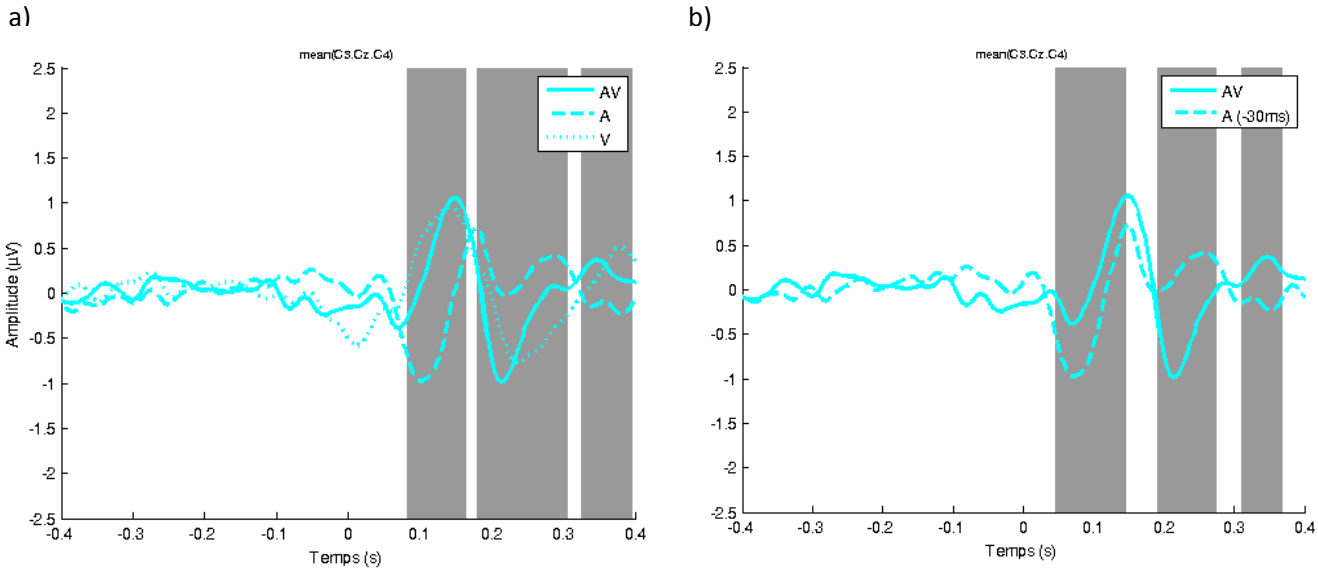

Figure 26. a) Potentiel évoqués lors de la présentation audiovisuelle, auditive, et visuelle pour les stimuli natifs contrôles. b) Différences observées lors du réalignement de la N1/P2 obtenues suite au décalage de la courbe auditive de - 30ms. Les zones grisées représentent les différences entre la modalité audiovisuelle et auditive ($p < .05$)

Nous pouvons donc conclure à une diminution d'amplitude de la N1 également dans le cas de la perception audiovisuelle de phonèmes natifs contrôles.

Nous constatons également que lors de la présentation audiovisuelle, l'onde P2 est réduite uniquement lors de la présentation du phonème /θ/. Dans les deux autres cas, son amplitude tend à être plus importante lors de la présentation audiovisuelle.

Enfin, en termes de latence, nous observons que deux des patterns ne présentent pas l'avantage généralement obtenu lors de la présentation visuelle. Seule la présentation audiovisuelle du phonème /s/ semble provoquer 30 ms d'avance sur les traitements auditifs.

3.10.2.1.2 P50

Lors de la présentation audiovisuelle du stimulus non natif /θ/, nous observons également une onde positive autour de 50 ms après l'*onset* auditif du stimulus. Nous supposons l'existence d'une P50 pour ce stimulus car celui-ci n'est pas connu du participant. Cette onde ne semble pas présente chez les hispanophones. Afin de le vérifier, nous avons effectué des t-tests continus contre la *baseline* (Figure 29).

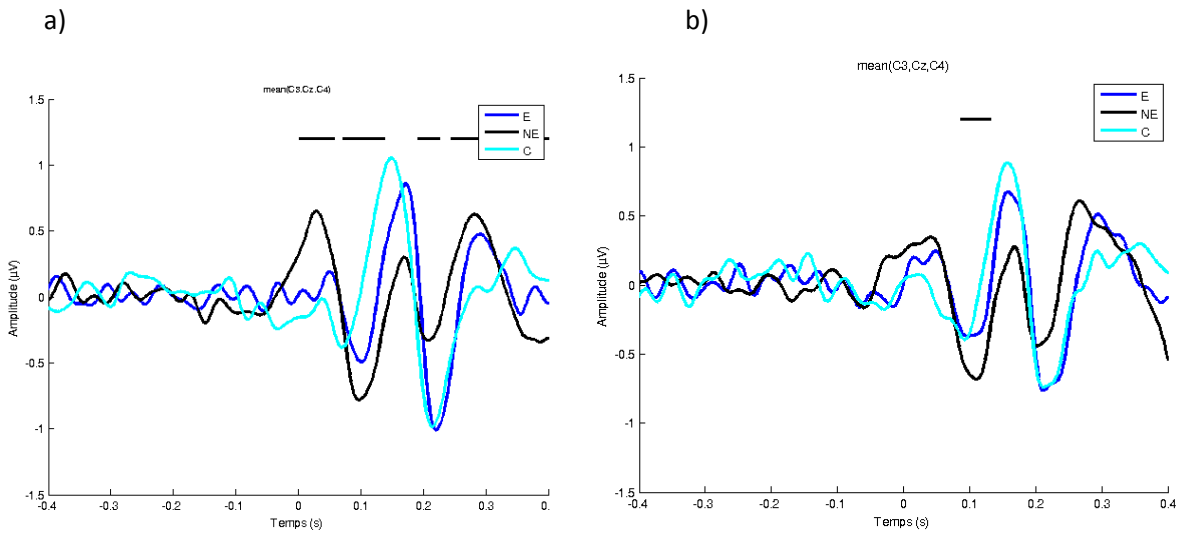

Figure 27. Comparaison de la courbe pour le stimulus non natif (noire) à la *baseline* pour les francophones (a) et les hispanophones (b). Les barres noires représentent les portions de signal pour lesquelles les activations générées par le stimulus NE sont significativement différentes de la baseline.

Les comparaisons à la *baseline* révèlent l'existence d'une P50 chez les francophones uniquement pour le stimulus non natif. En revanche, celle-ci n'est pas évoquée chez les hispanophones.

3.10.2.1.3 HISPANOPHONES

Pour les hispanophones, les stimuli contenant les phonèmes /f/ et /s/ induisent eux aussi une diminution de l'amplitude lorsqu'ils sont présentés en modalité audiovisuelle par rapport à la présentation auditive (Figure 30 et 31 a). Concernant l'amplitude de P2, il semblerait que pour cette population, tout comme pour les francophones, celle-ci ait tendance à être plus ample lors de la présentation audiovisuelle.

Concernant les variations de latence, et malgré le réalignement des pics par décalage de la courbe auditive (Figure 30 et 31 b), nous ne pouvons pas conclure à une différence significative puisque cette différence s'observe de concert avec des variations d'amplitude.

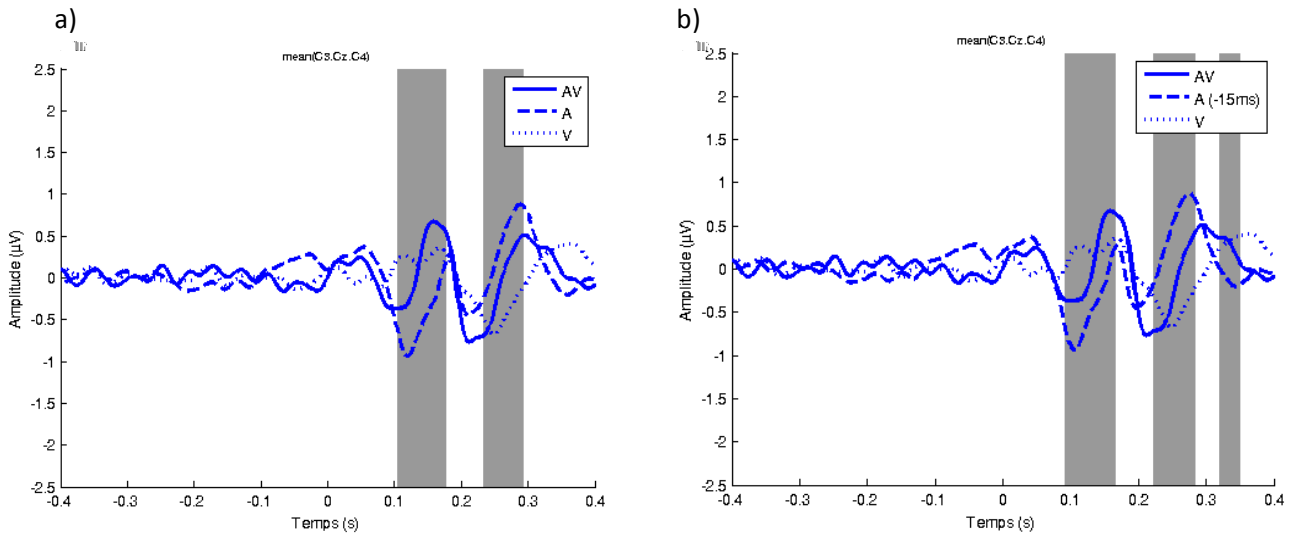

Figure 28. a) Potentiel évoqués lors de la présentation audiovisuelle, auditive, et visuelle pour les stimuli natifs. b) Différences observées lors du réalignement de la N1/P2 obtenues suite au décalage de la courbe auditive de 15 ms. Les zones grisées représentent les différences entre la modalité audiovisuelle et auditive ($p < .05$)

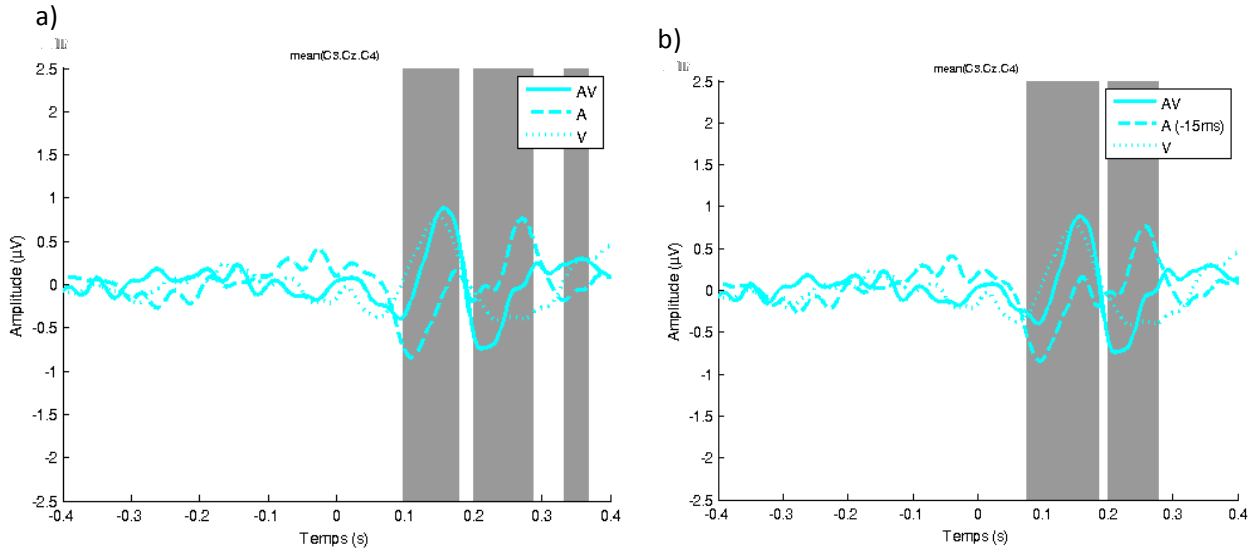

Figure 29. a) Potentiel évoqués lors de la présentation audiovisuelle, auditive, et visuelle pour les stimuli natifs contrôles. b) Différences observées lors du réalignement de la N1/P2 obtenues suite au décalage de la courbe auditive de 15 ms. Les zones grisées représentent les différences entre la modalité audiovisuelle et auditive ($p < .05$)

Enfin nous observons que la perception audiovisuelle du phonème non natif n'induit pas de modulation de N1 ou de P2, que ce soit en terme de latence ou d'amplitude (Figure 32).

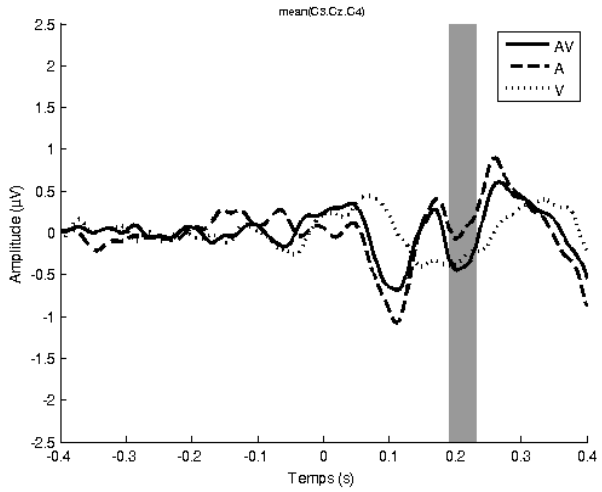

Figure 30. Potentiels évoqués lors des présentations auditive, audiovisuelle et visuelle chez les hispanophones pour les stimuli non natifs. Les zones grisées représentent les différences entre la modalité audiovisuelle et auditive ($p < .05$)

La comparaison directe des deux groupes de participants ne montre pas non plus de différence dans le traitement des stimuli non natifs, qu'ils soient présentés en modalité auditive, audiovisuelle ou visuelle (Figure 33 a, b et c respectivement).

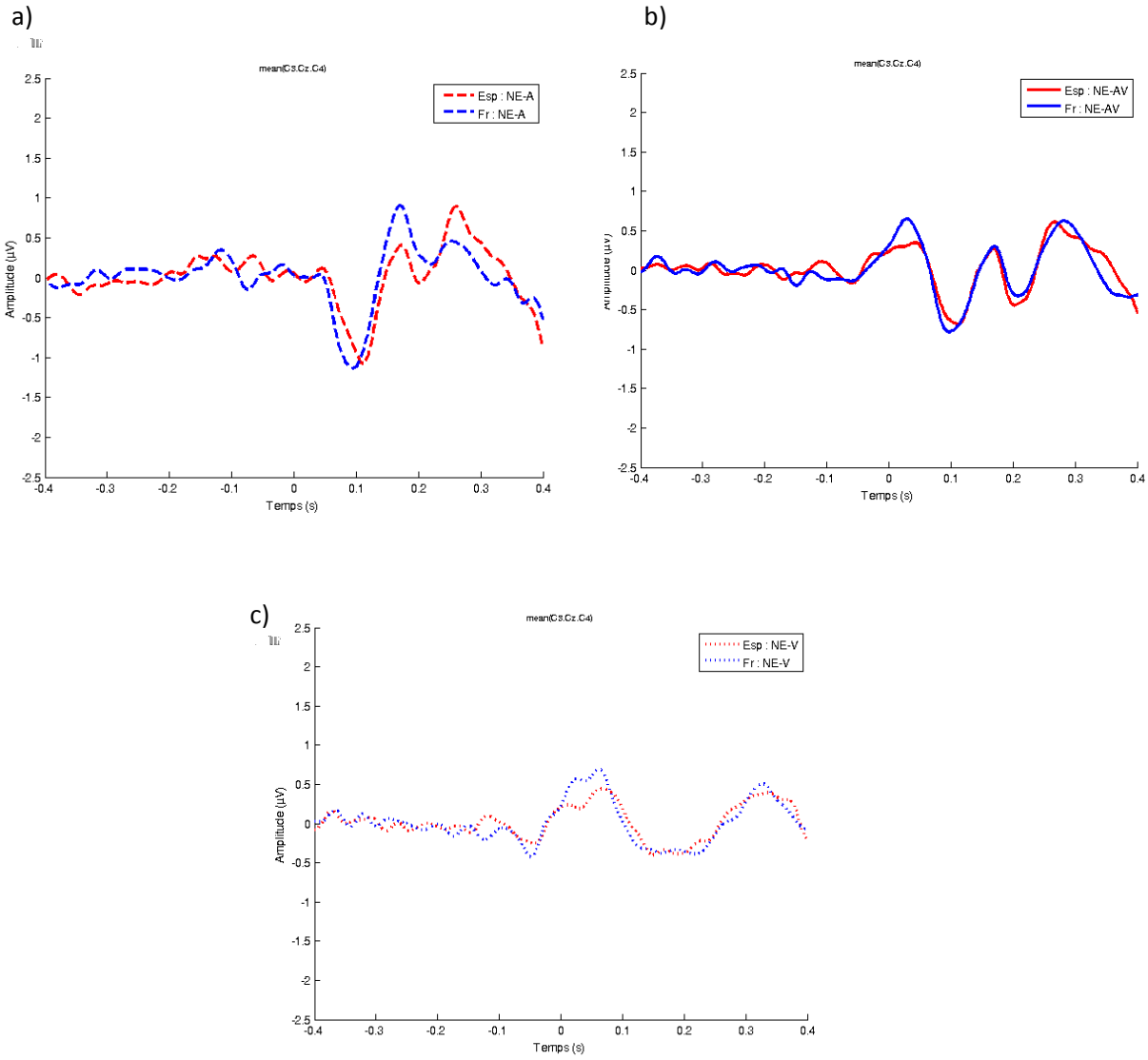

Figure 31. Potentiels évoqués lors des présentations auditive (a), audiovisuelle (b) et visuelle (c) chez les francophones (en bleu) et les hispanophones (en rouge) pour les stimuli non natifs. Les zones grisées représentent les différences entre la modalité audiovisuelle et auditive ($p < .05$)

3.11 DISCUSSION

Dans l'étude 1 présentée au chapitre 3, nous avons vu que les informations visuelles sur les gestes articulatoires du locuteur permettent de surmonter la surdité phonologique. Nous souhaitions reproduire ces résultats et mieux comprendre les processus neuronaux mis en jeu lors de la perception audiovisuelle de phonèmes non natifs. van Wassenhove et al. (2005) ont mis en évidence que la présentation audiovisuelle de phonèmes natifs produit une réduction de l'amplitude et de la latence des potentiels évoqués auditifs par rapport à une présentation auditive seule. Le but de cette étude était d'examiner l'effet de l'information visuelle sur les composantes auditives précoces lors des processus d'identification de phonème non natifs. Les participants, francophones et hispanophones natifs, devaient dans un premier temps mémoriser une séquence CV (où V = /a/) de référence contenant une consonne qui existe dans leur répertoire phonologique (i.e., /f/). Par la suite, trois types de stimuli leur étaient présentés. Ils différaient sur la consonne. La syllabe pouvait contenir le même phonème consonantique que la référence (i.e., /f/), ou un phonème qui n'existe pas dans le répertoire phonologique des francophones mais qui existe dans celui des hispanophones (i.e., /θ/), ou enfin un phonème existant dans les deux langues qui servait comme contrôle (i.e., /s/). Les participants devaient indiquer si la syllabe qu'ils entendaient, voyaient, ou voyaient et entendaient à chaque essai était la même que la référence. Nous avons enregistré les signaux EEG ainsi que les réponses comportementales des deux groupes de participants.

3.11.1 COMPORTEMENT

Conformément à nos attentes, la présentation auditive des stimuli non natifs a mis les participants francophones en difficulté. Nous observons une surdité phonologique lors de la perception auditive du phonème non natif /θ/. Ils obtiennent un score moyen de 57% de réponses correctes contre 98% (SD = 0.9) et 98% (SD= 0.5) pour les stimuli natifs et natifs contrôle respectivement. Lorsque la syllabe /θa/ est présentée, ils n'arrivent pas à la distinguer de la référence (i.e., /fa/) dans presque un cas sur deux. Cela reflète un phénomène d'assimilation phonologique. Ainsi, quand ce phonème non natif est perçu, il est assimilé à une catégorie existante du répertoire phonologique, dans ce cas /f/. Ce pattern de réponse est lié au fait que ce phonème n'existe pas dans l'inventaire phonologique des francophones. En effet, les hispanophones n'ont aucune difficulté à répondre que les séquences /θa/ sont

différentes de la séquence de référence /fa/.

Comme c'était le cas dans l'étude 1, les informations visuelles améliorent les capacités de discrimination des participants. Nous avons en effet obtenu un effet principal de la modalité de présentation avec un avantage lors de la présentation audiovisuelle. Le gain global permis par la présentation audiovisuelle est de 13,5% pour les francophones. Nous avons donc répliqué des résultats qui sont cohérents avec ceux de Hardison (2003) et Wang et al. (2008) qui obtenaient des améliorations de la discrimination lors de la présentation visuelle pour des visèmes qui n'existent pas dans le répertoire phonologique natif (i.e., le visème de l'interdentale /θ/ pour les coréens). Malgré les variations dans les réalisations articulatoires spécifiques à chaque langue, il semblerait que les informations visuelles permettent d'améliorer la discrimination des syllabes. Même lorsque les mouvements articulatoires ne sont pas familiers pour le participant, mais qu'ils sont assez saillants, comme c'est le cas de l'interdentale, ils permettent une amélioration des scores, dans notre cas d'environ 40% pour les phonèmes les plus difficiles à percevoir lors de la présentation auditive. Cela permet de ramener les performances des francophones au niveau de celles des hispanophones, avec des scores respectifs en présentation audiovisuelle de 97 et 98%. Nos résultats vont en revanche à l'encontre du modèle de Hazan et al. (2006) qui postule que les mouvements articulatoires ne peuvent être utilisés afin de désambigüiser une contraste non natif que lorsque ceux-ci sont partagés par les deux langues (e.g., le visème de /v/ peut être utilisé par les hispanophones pour distinguer /b/-/v/ puisque le visème labiodental existe en espagnol).

Les comparaisons entre les deux groupes de participants confirment également nos hypothèses, puisque la seule différence qui existe entre eux est observée pour le phonème /θ/ perçu auditivement. Les variations des réalisations acoustiques inter-langues d'un phonème donné ne sont donc pas une difficulté pour le locuteur non natif du moment que le phonème est connu du locuteur (i.e., /f/ et /s/). Lorsque celui-ci est inconnu (i.e., /θ/), l'information auditive ne sera pas suffisante pour discriminer les phonèmes, mais les informations visuelles sur les mouvements articulatoires du locuteur pourront compenser cette difficulté.

Nous avons également observé que chez les francophones, lors de la présentation audiovisuelle, les scores obtenus pour les stimuli non natifs /θ/ restent moins importants que ceux des stimuli natifs contrôles /s/. En effet, malgré l'apport des informations visuelles, les participants ont toujours plus de difficultés pour répondre que /θ/ est différent de /f/ lors de la présentation audiovisuelle alors que cette distinction est aisée lors de la présentation de /s/. Ce sont ici les caractéristiques acoustiques des phonèmes natifs qui permettent le plus cette

distinction car nous pouvons observer l'effet inverse lors de la présentation visuelle. En effet, lorsque seules les informations visuelles sont disponibles, /θ/ produit plus de réponses correctes que /f/ et /s/. Lors de la présentation visuelle de /fa/ ou /sa/ les participants sont moins à même de répondre que le premier est le même stimulus que sa référence, et pour le second de répondre qu'il est différent. Ce pattern différentiel (/f/ = /s/ > /θ/ en audiovisuel et /f /= /s/ < /θ/ en visuel) peux s'expliquer par le fait que lors de la présentation audiovisuelle, les réponses sont guidées préférentiellement par la modalité auditive pour les phonèmes natifs (puisque cette information est suffisante pour répondre qu'il n'y a pas de différence entre ces phonèmes lors de la présentation auditive). C'est également le cas pour les phonèmes non natifs, ce qui se traduit, pour ce dernier en une gêne de la prise en compte des informations auditives se traduisant par des performances inférieures lors de la présentation audiovisuelle par rapport à visuelle seule.

L'observation des réponses des espagnols nous indique que le phonème /f/ porte à confusion. Un pattern assez similaire à celui des francophones est globalement obtenu. Les scores de la présentation visuelle de /f/ sont plus faibles que ceux obtenu dans les deux autres modalités et nous observons également des scores plus faibles pour /f/ par rapport aux deux autres phonèmes lors de la présentation visuelle. Il semblerait donc que, pour les hispanophones, les caractéristiques articulatoires de /s/ et /θ/ soient plus discriminantes que /f/.

3.11.2 NEUROPHYSIOLOGIE

Les analyses EEG ont révélé de nombreux points intéressants.

3.11.2.1 IMPACT DE LA PRESENTATION AUDIOVISUELLE SUR L'AMPLITUDE DES PE AUDITIFS

Nous avions formulé l'hypothèse d'une diminution de l'amplitude de la composante N1/P2 lors de la présentation audiovisuelle par rapport à la présentation auditive. Une réduction de l'amplitude de N1 a été observée pour tous les stimuli chez les francophones. Cela est en accord avec les résultats de Besle et al. (2004), Pilling (2009) ou van Wassenhove et al. (2005) qui obtenaient une réduction de l'amplitude de la N1 lors de la présentation audiovisuelle. Puisque les trois types de stimuli sont sujets à cette réduction d'amplitude, nous pouvons statuer sur le fait qu'un mouvement articulatoire, même inconnu, permet de moduler la réponse corticale auditive. Cela est en accord avec les résultats de Stekelenburg & Vroomen (2007) qui observaient que la N1 était uniquement sensible au fait que des mouvements anticipatoires soient présentés avant l'entrée auditive, même si ceux-ci étaient incongruents. Par exemple, qu'un /fu/ auditif soit combiné avec un /bi/ visuel ou que le bruit d'un claquement de main soit généré par une cuillère qui frappe une tasse, l'amplitude de la N1 était réduite lors de la présentation audiovisuelle par rapport à la présentation auditive. Cette onde n'est donc pas sensible au contenu de la vidéo en lui-même, mais simplement au fait qu'une entrée auditive puisse être anticipée par des mouvements qui indiquent qu'une « entrée auditive » va arriver. Il en va donc de même pour les mouvements articulatoires, qu'ils soient connus ou inconnus. Cette réduction reflète une facilitation des traitements des indices acoustiques lorsqu'ils sont amorcés par de l'information visuelle. L'information visuelle permet en effet de réduire l'incertitude du système ainsi que les demandes computationnelles du cortex auditif (Besle et al., 2004; van Wassenhove et al., 2005).

Toutefois, les hispanophones ne présentent pas de modulations de la N1 lors de la présentation audiovisuelle de /θ/. L'explication la plus probable est en rapport avec le manque de puissance lié au faible nombre de participants de ce groupe puisque seuls 13 participants ont été testés. Dans ce sens, nous pouvons d'ailleurs constater que chez les hispanophones, la comparaison entre les potentiels générés lors de la présentation de phonèmes non natifs et la *baseline* (effectuée afin de mettre à jour la P50 chez les francophones, Figure 29 a) indique seulement une différence sur la N1 (celle-ci étant significativement différente de la *baseline*,).

Cela souligne encore une fois le manque de puissance de ce groupe dont l'effectif n'est pas suffisant. De fait, nous observons une différence tendancielle, même si celle-ci n'est pas significative.

Nous obtenons également un pattern différentiel entre N1 et P2. Alors que des modulations conjointes de N1/P2 sont généralement observées, ce n'est pas le cas dans notre étude (à l'exception des stimuli non natifs perçus par les francophones). Les modulations observées sur la P2 (aussi bien en fonction du phonème que de la modalité de présentation) sont beaucoup plus inconsistantes que celles observées sur la N1. Nous avons observé :

(1) Un effet inverse (amplitude P2 AV > A) sur /s/ pour les francophones et sur /f/ et /s/ chez les espagnols ;
(2) Absence de modulation sur /f/ chez les francophones et /θ/ chez les hispanophones (notons que N1 ne subit également aucune modulation) ;
(3) Effet cohérent (amplitude P2 AV < A) pour /θ/ chez les francophones.

Devant cette hétérogénéité, il est difficile de conclure quant aux raisons de telles modulations. En effet, une réduction des composantes lors de la présentation audiovisuelle est régulièrement obtenue dans la littérature (Arnal et al., 2009 ; Besle et al., 2004 ; Klucharev et al., 2003 ; Stekelenburg & Vroomen, 2007 ; van Wassenhove et al., 2005). Elle est même observée dans le cadre des stimuli non-langagiers, du moment que des mouvements préparatoires sont disponibles (Stekelenburg & Vroomen, 2007). Ceci dit, nous ne savons pas s'il y avait dans nos stimuli des informations visuelles sur les gestes du locuteur assez saillantes pour être détectées, car nous n'avons pas réalisé des mesures articulatoires. Il est possible que cette hétérogénéité au niveau de la P2 soit due au fait qu'il n'y ait pour certains stimuli que peu d'anticipation (cf. Schwartz & Savariaux, 2014), dans le cas de /f/ par exemple. Cependant, les modulations observées sur N1 semblent montrer que, même si les mouvements préparatoires sont faibles, ils sont bien présents et traités.

Nous pouvons cependant questionner le manque de puissance observé, celui-ci peut être dû, soit à une intensité sonore trop faible lors des passations, soit à la nature des stimuli auditifs utilisés. L'expérimentation s'est déroulée dans un environnement calme, partiellement insonorisé, et le volume des stimuli, était à un niveau confortable et maintenu constant pour tous les participants. Il est donc peu probable que les faibles amplitudes soient liées à l'intensité sonore trop faible. Concernant la nature des stimuli, les fricatives bénéficient d'un *onset* auditif plus distribué et dont l'intensité augmente graduellement durant les premières

millisecondes par rapport par exemple aux plosives utilisées par van Wassenhove et al. (2005). « L'augmentation de l'énergie étant graduelle après le silence, il est souvent difficile de déterminer exactement le début de la fricative en position initiale absolue » (Ridouane, 2003, p.56). En effet, le temps de relâchement de la friction est plus important, et celle-ci atteint donc son amplitude maximum plus lentement que les plosives (Johnson, 2003 ; Kluender & Walsh, 1992 ; Walsh & Diehl, 1991) dont le relâchement du *burst* est soudain. De plus, alors qu'un *burst* est un signal de haute intensité, les fricatives non voisées articulées à l'avant du conduit vocal, comme les labiodentales ont une intensité relativement faible (Strevens, 1960). Cela peut expliquer la faible amplitude de nos signaux. Cependant, rappelons que nos données montrent une modulation de l'amplitude de N1 lors de la présentation audiovisuelle pour tous les stimuli à l'exception de /θ/ pour les hispanophones, il est donc peu probable que le matériel soit en lui même l'explication pour les différences observées sur la modulation de P2.

Il est à noter qu'une modulation conjointe des deux ondes n'est pas systématiquement observée. Certains auteurs ont observé une diminution uniquement de N1 (Treille et al., 2014) ou uniquement de P2 (Baart, Vroomen, et al., 2014 ; Miki, Watanabe, & Kakigi, 2004 ; Möttönen, Schürmann, & Sams, 2004). Il semblerait en effet que ces deux potentiels aient un fonctionnement distinct. Le premier refléterait des processus de bas niveau alors que le second permettrait notamment de distinguer entre des processus spécifiques liés à la parole et des processus plus généraux (observables dans des contextes non langagiers). La P2, contrairement à la N1, est par exemple insensible à la prédictibilité des informations auditives mais sensible à la congruence phonétique. Par exemple, entendre un /ba/ durant la production d'un /fu/ augmente la P2 par rapport à un stimuli congruent alors que la congruence n'a pas d'effet sur la N1 (Klucharev et al., 2003 ; Stekelenburg & Vroomen, 2007). Dans une étude récente, Baart, Stekelenburg et Vroomen (2014) ont montré que l'intégration audiovisuelle observée au travers de la réduction de P2 reflétait l'étape durant laquelle les informations auditive et visuelle sont liées à un niveau *phonétique*. En effet, la réduction de la P2 n'était observable que lorsque les participants (qui perçoivent tous de la parole synthétique sinusoïdale) étaient en mode « parole ». Il semblerait donc que pour une raison ou une autre, certains stimuli aient nécessité un traitement plus important lors de la présentation audiovisuelle, car plus qu'une absence de modulation de la P2, nous obtenons des résultats inverses à ceux obtenus précédemment dans la littérature (potentiels plus amples lors de la présentation audiovisuelle) pour trois conditions (/f/ pour les francophones et /s/ et /θ/ pour les hispanophones). Il semble peu probable que les caractéristiques des stimuli soient

responsables de l'inversion observée puisque les modulations d'amplitude sont différentes, à l'intérieur d'une même modalité pour les mêmes stimuli en fonction du groupe. Par exemple, alors que la présentation de /f/ en audiovisuel ne provoque aucune différence d'amplitude chez les francophones, les hispanophones ont un pattern d'inversion avec une augmentation de l'amplitude lors de la présentation audiovisuelle. La réponse à la question posée par ce pattern de résultats reste pour l'instant ouverte.

En ce qui concerne les modulations de la latence de N1, les réductions n'ont été observées que pour un phonème chez les francophones (i.e., /s/) et deux phonèmes chez les hispanophones (i.e., /s/ et /f/). Ces résultats ont soulevé de nombreuses questions, méthodologiques plus que théoriques. En effet, cette réduction de la latence s'observe généralement de concert à une réduction de *l'amplitude* de N1, induite par la bimodalité.
L'absence de modulation concernant le phonème /θ/ en terme de latence ou d'amplitude en fonction de la modalité de présentation chez les hispanophones peut cependant trouver sa cause dans les différences d'utilisation de ce phonème en espagnol.

Enfin, lorsque l'on compare les trois types de stimuli lors de la présentation audiovisuelle, nous avons observé une précocité des traitements réalisés sur /s/ pour les N1 et P2 uniquement chez les francophones alors que cela n'était le cas pour aucun autre des stimuli pour aucun des deux groupes. La difficulté d'interprétation de ce type de résultats est induite par le fait que des traitements précoces distincts soient obtenus pour un même phonème entre les deux populations, mais également pour deux phonèmes bénéficiant de caractéristiques assez similaires (/f/ est une fricative non voisée qui fait partie du répertoire phonologique des francophones, comme /s/). Ce phénomène est lié à la présentation des informations articulatoires puisqu'aucune différence de *latence* n'apparaît entre ces deux stimuli lors de la présentation auditive et que seule une faible différence de latence est obtenue entre les stimuli /f/ et /s/, au détriment de ce dernier lors de la présentions visuelle seule. Considérant ces observations, il nous est difficile d'envisager les raisons d'une différence de latence induite par ce phonème lors de la présentation audiovisuelle.

3.11.2.2 DETECTION PRE-ATTENTIVE D'EVENEMENTS VISUELS NOUVEAUX

Un élément intéressant de nos résultats concerne l'observation d'un P50 lors de la présentation audiovisuelle du phonème non natif. Elle n'est observée que lorsque les francophones perçoivent /θ/ en modalité audiovisuelle et n'est pas obtenue chez les

hispanophones. De plus, elle n'apparaît dans aucune des autres modalités de présentation. Il faut donc la considérer comme une modification des traitements auditifs générés uniquement lors de la présentation multimodale. Cette P50 apparaît entre 20 et 30 ms.

Un pan de la littérature considère cette onde comme un marqueur de filtrage sensoriel. Cette onde serait un filtre qui protégerait les centres corticaux de haut niveau d'être débordés par des informations non pertinentes. Un phénomène de « *gate out* » ou « *filter out* » se mettrait en place suite à l'apparition d'informations redondantes. Ce phénomène permettrait de ne pas laisser passer des informations redondantes ce qui se traduirait par une réduction de l'amplitude de la P50. Ce phénomène de filtrage permettrait d'alléger les traitements successifs suite à l'apparition d'information déjà « connue » du système. Par exemple, lorsque deux clics successifs sont présentés, l'onde de la P50 pour le second est moindre que pour le premier (Freedman, Adler, Waldo, Pachtman, & Franks, 1983). Grunwald et al. (2003) ont mis en évidence chez huit sujets des ondes P50 répondant au premier click (ligne continue, Figure 34) et des modulations significativement moins importantes pour le second click (ligne pointillée).

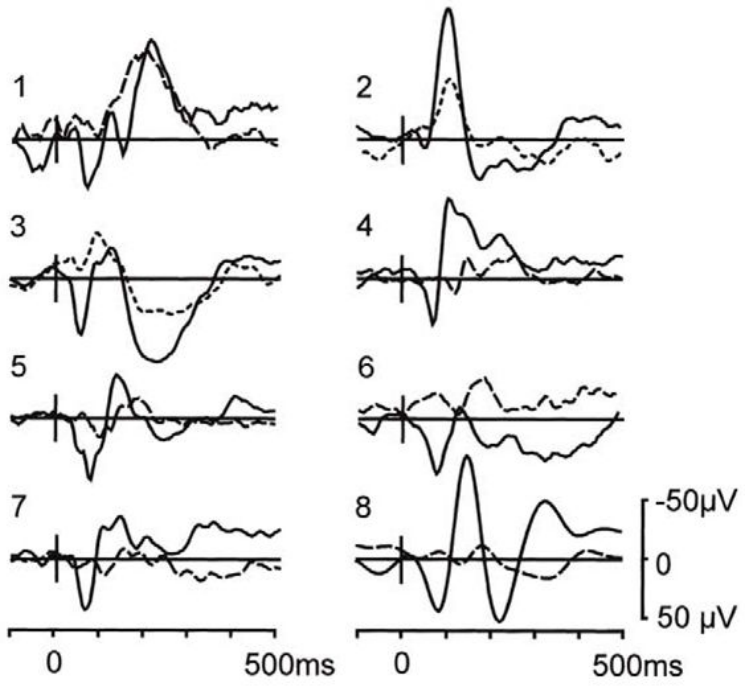

Figure 32. Modulations de la P50 lors de la présentation d'un premier clic (ligne continue) et d'un second clic (ligne pointillée). (Tiré de Grunwald et al., 2003).

A l'inverse, le phénomène de « *gate in* » permettrait de faciliter le passage de informations nouvelles ou incongruentes (donc non redondantes) qui se traduit par une augmentation de l'amplitude de la P50 (Boutros et al., 1995). L'étude de Lebib et al. (2003) observe cette onde lors de la perception de la parole audiovisuelle. Ils voulaient, via l'observation de la P50, fournir des données montrant que le cerveau peut détecter des changements lors de la perception bimodale à un niveau pré-attentif (ou à une étape très précoce) et montrer que la détection de ces changements dépend de la congruence et du niveau de discrimination de la parole audiovisuelle. Leur stimuli étaient des voyelles isolées (/i/, /a/, /ø/ et /y/) qui pouvaient être congruentes (C), ou incongruentes (I) au regard de l'articulation, mais également divisées en catégories : faciles à distinguer (E) ou difficiles (H) à distinguer. Les participants sont donc testés dans 4 conditions expérimentales : CE, CH, IE, IH. Le participant effectuait un jugement de congruence. Leurs résultats montrent deux effets. Lors des comparaisons **C**E vs **I**E et **C**H vs **I**H (c'est-à-dire lors de l'observation de l'effet de la congruence), ils n'ont obtenu aucune différence entre les deux conditions lorsque les stimuli étaient difficiles à distinguer. Cependant, lorsque les stimuli étaient faciles à distinguer, le pic de la P50 était plus important pour les entrées incongruentes. Lorsqu'ils ne s'intéressaient qu'à l'effet de la discriminabilité (comparaisons CE vs CH et IE vs IH) aucun effet n'était observé pour les stimuli congruents, alors que lors de la comparaison IE vs IH, IE était associé à une onde plus importante. Les deux résultats convergent donc et indiquent que la P50 semble plus importante lorsque les entrées perçues sont facilement interprétables comme étant déviantes. Lebib et al. (2003) interprètent ces résultats comme une détection cérébrale précoce d'informations audiovisuelles non redondantes. Les informations visuelles modulent donc la détection pré-attentive de stimuli bimodaux incongruents ou non pertinents. Nous pouvons donc ajouter que cette « fonctionnalité » de la P50 est également effective lors de la présentation multimodale.

Ces observations sont cohérentes avec nos résultats. En effet, dans notre cas, tous les stimuli sont congruents et nous n'observons de P50 que lors de la présentation du phonème non natif chez les francophones. Elle pourrait donc refléter un filtrage sensoriel induit par les informations visuelles. En effet, alors que l'absence de celle-ci dans les autres conditions audiovisuelles nous indique que le système a perçu une entrée visuelle qui correspondait à l'entrée auditive, l'apparition de cette onde lors de la perception de /θ/ peut alors s'expliquer par une détection pré-attentive d'un signal auditif que le système considère comme non redondant aux informations visuelles perçues. Soit cela est dû à la nature du signal visuel, qui

est inconnu et qui par conséquent n'a pas lieu d'être « filtré » puisque le signal auditif qui va suivre est, pour le coup, incertain. La détection de mouvements préparatoires dont ont ne connaît pas les conséquences engendrerait donc une P50. Soit cela est dû au fait que le système considère que les signaux visuels et auditifs sont dans ce cas incongruents. Le filtre sensoriel est donc « levé » lors de l'apparition des informations auditives sur le stimulus qui est alors considéré comme une information nouvelle. Cette onde ne serait alors pas observée lors de la présentation visuelle puisque aucune information auditive n'apparait consécutivement à l'information visuelle. Ainsi, aucune modulation de la P50 serait générée puisque ce filtrage n'est activé que lors d'une double information distribuée dans le temps (auditif -auditif ; visuelle - auditif).

3.12 CONCLUSION GENERALE

Notre étude a dans un premier temps permis de répliquer au niveau comportemental le phénomène de surdité phonologique chez les francophones en modalité auditive seule ainsi que l'impact positif qu'entraîne la présentation audiovisuelle sur ce phénomène. Au niveau des données neurophysiologiques, même si celles-ci présentent des divergences avec celles obtenues précédemment dans la littérature, notre étude amène quelques éléments nouveaux.

Nous avons obtenu des patterns différentiels entre N1 et P2 en termes de modulation de l'amplitude induite par la présentation audiovisuelle. La première est réduite lors de la présentation audiovisuelle pour presque tous les phonèmes (à l'exception de /θ/ chez les hispanophones, ce qui peut être expliqué par un manque de puissance statistique lié au faible nombre de participants). Ceci est cohérent avec le pattern de résultats d'études précédentes mettant en évidence que les informations visuelles sur les gestes articulatoires du locuteur permettent d'alléger les traitements effectués par le cortex auditif (Arnal et al., 2009 ; Besle et al., 2004; Klucharev et al., 2003 ; Pilling, 2009 ; Stekelenburg & Vroomen, 2007 ; van Wassenhove et al., 2005). Ces observations ont été faites avec des stimuli présentant des plosives, notre travail étend ces résultats à des signaux de parole utilisant des consonnes fricatives, qui sont acoustiquement moins stables. En effet, les fricatives disposent d'indices acoustiques plus distribués et notamment d'une augmentation graduelle de l'amplitude du signal qui aurait potentiellement pu limiter l'observation de la modulation induite par la modalité de présentation. Notons toutefois que cet effet n'est pas spécifique au langage et dépend, selon certains auteurs, de la présence d'un mouvement préparatoire avant le début du

signal acoustique (Stekelenburg & Vroomen, 2007).

La modalité de présentation a des effets inconsistants au regard de la littérature, aussi bien en fonction du Groupe que du Type de stimuli. Des patterns inverses à ceux de la littérature ont été observés lors de la présentation de /s/ chez les francophones ainsi que /f/ et /s/ chez les hispanophones. Une absence de modulation a été observée pour /f/ chez les francophones et pour /θ/ chez les hispanophones. Enfin, un pattern en accord avec ceux obtenus dans la littérature n'est obtenu que durant la présentation de /θ/ pour les francophones. Nous ne sommes pas en mesure d'expliquer ce pattern.

La contribution la plus importante de ce travail est que nos données indiquent une différence de traitement liée à l'existence ou non d'un phonème dans la langue maternelle. Nous avons observé une modulation de N1 lors de la présentation audiovisuelle chez les francophones lorsqu'ils percevaient /θ/. Nous avons interprété cet effet comme lié à l'apparition d'une P50. En effet, une P50 a été observée uniquement lors de la présentation audiovisuelle d'un phonème inconnu (aucune P50 n'était observée chez les hispanophones). Elle suggère que /θ/ est considéré comme nouveau et/ou non redondant. L'apparition de la P50 augmenterait par la suite le traitement auditif des caractéristiques acoustiques du signal, ce qui se matérialiserait par une augmentation de l'amplitude de la N1 pour ce phonème par rapport aux phonèmes connus (e.g., /f/ pour les francophones ou /θ/ pour les hispanophones). Il semblerait donc que la perception audiovisuelle de phonèmes non natifs module les traitements pré-attentifs au niveau de la P50.

3.13 RESUME

But de l'étude : Déterminer les processus neurophysiologiques mis en place lors de la perception audiovisuelle de phonèmes natifs et non natifs.

Populations : 19 francophones natifs et 13 hispanophones natifs testés à Grenoble.

Protocole : Les stimuli contenaient soit un phonème qui existe en français (i.e.,/f/), un phonème qui n'existe pas en français (i.e., /θ/), ou un distracteur qui existe en français (/s/). Les consonnes étaient insérées dans des séquences monosyllabiques en contexte vocalique /a/. Le paradigme consistait en la présentation d'un « mot » de référence qui devait être mémorisé au début du bloc, puis être comparé aux « mots » présentés par la suite. Pour chacun de ces

stimuli, le participant devait répondre si ce « mot » était le même ou un mot différent du mot de référence. Le pourcentage de réponses correctes ainsi que l'activité neuronale ont été enregistrés.

Résultats : Les résultats comportementaux révèlent un phénomène de surdité phonologique qui se met en place lors de la perception auditive des phonèmes non natifs par les francophones. Cette surdité est largement surmontée lors de la présentation audiovisuelle. Les résultats neurophysiologiques ont mis en évidence une réduction de l'amplitude de N1 lors de la présentation audiovisuelle (ce qui est cohérent avec un effet d'amorçage des informations visuelles, qui permettent donc d'alléger les traitements effectués par le cortex auditif (Arnal et al., 2009 ; Besle et al., 2004 ; Klucharev et al., 2003 ; Pilling, 2009 ; Stekelenburg & Vroomen, 2007 ; van Wassenhove et al., 2005), sans toutefois engendrer de réduction de la latence. Lors de la présentation audiovisuelle, l'amplitude de la P2 était différente selon le type de stimulus présenté et de la modalité. L'inconsistance de ces variations est difficile à interpréter. Les analyses N1 et P2 au niveau temporel durant la présentation audiovisuelle n'ont pas donné lieu à une diminution de la latence. Enfin, une P50 a été observée uniquement lors de la présentation audiovisuelle d'un phonème inconnu. De plus, la P50 n'a pas été observée chez les hispanophones. Cela suggère que ce phonème inconnu est considéré comme nouveau et/ou non redondant, ce qui module tout le traitement auditif subséquent.

Conclusion : Nos résultats répliquent le phénomène de surdité phonologique pour des phonèmes rarement utilisés dans la littérature. La génération d'une P50 lors de la présentation audiovisuelle de phonèmes inconnus devra être soumise à de plus ample investigation afin de déterminer la nature de ce processus et son rôle dans la perception de la parole bimodal

CHAPITRE 5

ETUDE 3

MODULATION DE LA SURDITE PHONOLOGIQUE EN FONCTION DE L'EXPERIENCE LINGUISTIQUE

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 21 October 2014
doi: 10.3389/fpsyg.2014.01179

Bilingualism affects audiovisual phoneme identification

Sabine Burfin[1], Olivier Pascalis[1], Elisa Ruiz Tada[2], Albert Costa[2,3], Christophe Savariaux[4] and Sonia Kandel[1,4,5]*

[1] *LPNC (CNRS UMR 5105) – Université Grenoble Alpes, Grenoble, France*
[2] *Universitat Pompeu Fabra, Barcelona, Spain*
[3] *Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain*
[4] *GIPSA-lab (CNRS UMR 5216) – Université Grenoble Alpes, Grenoble, France*
[5] *Institut Universitaire de France*

Article complet disponible en Annexe 2

4.1 INTRODUCTION

Comme nous l'avons vu dans le Chapitre 1, le phénomène de surdité phonologique se met en place précocement dans le développement. Il est dû à l'affinage perceptif qui se met en place suite à l'exposition répétée à notre langue maternelle. A la naissance, les enfants sont en effet capables de discriminer tous les phonèmes de toutes les langues (Kuhl et al., 2006 ; Werker & Tees, 1984). Cette capacité diminue progressivement durant la première année de vie pour la plupart des phonèmes qui n'existent pas dans notre langue (Mattock & Burnham, 2006 ; Mattock, Molnar, Polka, & Burnham, 2008 ; Werker, Gilbert, Humphrey & Tees, 1981 ; Werker & Tees, 1984). Cependant le phénomène de surdité phonologique peut se réduire via l'utilisation des mouvements articulatoires visibles. Les études 1 et 2 ont en effet montré que des indices visuels améliorent l'identification des contrastes non natifs (pour d'autres études voir : Davis & Kim, 2004 ; Erdener & Burnham, 2005 ; Hardison, 1999, 2003, 2005 ; Hazan et al., 2006 ; Kluge, Reis, Nobre-oliveira & Bettoni-techio, 2006 ; Sekiyama, Kanno, Miura & Sugita, 2003 ; Thompson & Hazan, 2007 ; Wang, Behne & Jiang, 2008, 2009). Cet avantage a été observé la plupart du temps avec des populations qui avaient des connaissances déjà avancées dans la langue étrangère (L2). Or, nous savons aujourd'hui que la sensibilité aux informations visuelles dépend de nombreux facteurs (cf. Partie 3.2.3. « Facteurs qui modulent l'utilisation des informations visuelles de la langue étrangère »), notamment la structure du répertoire phonologique. L'exposition précoce à une langue donnée, ou à plusieurs langues semble également impacter l'utilisation des informations visuelles de langue inconnue, au moins durant la petite enfance (Weikum et al., 2007). Mais est-ce que le fait de posséder plusieurs répertoires phonologiques modulerait l'utilisation des indices visuels lors de la perception de phonèmes inconnus ? Afin de répondre à cette question, nous questionnerons la capacité d'une population de bilingues à discriminer un contraste inconnu, sur la base des informations auditive et audiovisuelle.

4.1.1 LE BILINGUISME ? NON ! LES BILINGUISMES

Avant de nous interroger sur ces processus, nous souhaitons attirer l'attention sur un point fondamental qui est à la base de la problématique de ce chapitre : la définition du bilinguisme. Selon l'étude de Romaine (1995), la moitié de la population du monde est bilingue. L'augmentation de la population multilingue depuis quelque années, notamment avec le

processus de globalisation, a facilité l'accumulation de connaissances sur le multilinguisme. En effet, David Crystal approximait en 2003 que les deux tiers des enfants du monde grandissaient dans un environnement plurilingue. Ne considérant que la population anglophone, 41% de la population est plurilingue en anglais et une autre langue soit plus de 235 millions de personnes. En Europe, Tabouret-Keller (2004) estimait que 50 % de la population étaient bilingue. Or, le terme « bilingue » est difficile à définir, tout comme le niveau de maîtrise d'une langue est une question délicate à traiter. Quelle est la différence entre le terme bilinguisme et la notion d'apprentissage d'une langue étrangère ? Lors du Donostia Workshop on Neurobilingualism (septembre 2010), les spécialistes dans le domaine se sont mis d'accord sur l'idée que le bilinguisme est impossible à définir et que les individus dits « bilingues » participant dans les études scientifiques doivent être caractérisés. Dans ce chapitre, nous caractériserons donc les participants de manière à spécifier rigoureusement leurs compétences linguistiques. Néanmoins, pour des raisons de simplicité, nous classerons les participants selon trois grandes catégories. La première comprendra des bilingues dits « de naissance »,« simultanés » ou «précoces» lorsque les individus ont été en contact avec deux langues depuis la naissance ou avant six ans. Dans cette catégorie, deux sous-types seront considérés : 1) les bilingues dits « équilibrés » qui ont une maîtrise équivalente dans les deux langues (Navarra, Sebastián-Gallés & Soto-Faraco, 2005) et 2) les bilingues dits « relatifs » qui ont une langue dite « dominante » parce qu'ils la maîtrisent mieux que l'autre (Marian, 2008). Le 2ème groupe comprendra des « bilingues successifs » ou « tardifs » lorsque les individus auront appris la langue étrangère après six ans. Enfin, la 3ème catégorie comprendra, quant à elle, des individus en cours d'apprentissage d'une langue étrangère lorsque l'apprentissage de la deuxième langue fait intervenir un cadre formel et un apprentissage explicite (l'école, par exemple). Chez ces derniers, établir une conversation dans une langue différente de la langue maternelle requiert une mobilisation importante de compétences, non seulement linguistiques, mais aussi attentionnelles et mnésiques.

La dichotomie « précoce/tardif » a une importance capitale. En effet, alors que les bilingues précoces ont un fonctionnement spécifique, les bilingues tardifs ont un fonctionnement plus similaire à celui des monolingues. Par exemple, les bilingues tardifs espagnol-anglais ont des frontières catégorielle de consonnes qui sont situées entre celles de l'espagnol et de l'anglais alors que celle des bilingues précoces sont similaires à celles des monolingues dans chacune de leur langue respectivement (Flege, 1995 cité dans Lee &

Iverson, 2011). Les bilingues tardifs ont également un accent plus marqué que les bilingues précoces (Fox, Flege & Munro, 1995). Ces deux groupes ont donc un fonctionnement distinct.

Nous allons dès à présent nous intéresser spécifiquement au premier groupe, qui se détache à la fois des bilingues tardifs et des monolingues. Nous allons décrire de manière succincte les modifications induites par l'exposition précoce à plusieurs langues.

4.1.2 QUELLES CONSEQUENCES ?

Le bilinguisme entraîne des modifications importantes dans les processus de traitement du langage : voir Adesope, Lavin, Thompson et Ungerleider, 2010 et Cook, 1997 pour des études sur la conscience métalinguistique ; Gollan et Acenas, 2004 pour une étude sur le phénomène de « *tip of the tongue* » ; Ransdell & Fischler, 1987 pour une étude sur les interférences en décision lexicale ou Byers-Heinlein, Burns et Werker, 2010) pour une étude sur les facultés des nouveau-nés bilingues à discriminer des langues. Le bilinguisme peut également avoir un impact dans des mécanismes aussi variées que le control cognitif (Bialystok, Craik & Luk, 2008 ; Bialystok, 2008 ; Costa, Hernández & Sebastián-Gallés, 2008), la catégorisation des concepts d'objets (Cook, Bassetti, Kasai, Sasaki & Takahashi, 2006), ou la perception des couleurs (Thierry, Athanasopoulos, Wiggett, Dering & Kuipers, 2009). Nous allons focaliser notre attention sur le traitement phonologique. En particulier, l'étude des bilingues de naissance permet d'obtenir des informations sur l'impact d'une familiarisation précoce à plusieurs codes phonologiques et donc oro-faciaux.

Des travaux indiquent que les bilingues auraient un fonctionnement phonologique spécifique, différent de celui des monolingues à plusieurs niveaux. Les bilingues semblent particulièrement sensibles aux différences qui existent entre les phonèmes, qu'ils soient natifs ou non natifs. Byers-Heinlein et al. (2010) présentent par exemple des données qui indiquent que les nouveau-nés ayant été exposés à une langue in-utero (i.e., mères de langue maternelle anglaise ou tagalog) préfèrent la langue de leur mère. Si des mères bilingues parlent les deux langues durant la grossesse, les nouveau-nés n'auront pas de préférence pour l'une ou l'autre des deux langues. De plus, si un mère bilingue parle anglais et chinois durant sa grossesse, le nouveau-né n'aura pas de préférence entre une langue parlée durant la grossesse (i.e., anglais) et une nouvelle langue (i.e., tagalog). Cela suggère que les nouveau-nés de milieux plurilingues traitent les sons de langage différemment de ceux qui naissent dans un

environnement monolingue (Burns, Werker, & Mcvie, 2003 ; Kuhl, Tsao & Liu, 2003). Cela se traduirait par des différences dans les structures neuronales du cortex auditif à l'âge l'adulte (Ressel et al., 2012).

Les bilingues semblent également tirer partie de l'information visuelle de parole. Dans ce sens, l'étude de Weikum et al. (2007) a apporté des données indiquant que l'expérience linguistique des enfants à la naissance est déterminante pour le développement de la sensibilité visuelle pour la discrimination des langues. Dans cette étude, des enfants de six et huit mois regardaient une vidéo silencieuse dans laquelle une bilingue franco-anglaise racontait une histoire soit en français, soit en anglais. Tous les enfants de six mois, qu'ils aient grandit dans un environnement monolingue anglophone ou dans un environnement bilingue franco-anglais, étaient capables de discriminer les deux langues visuellement. A huit mois, les enfants monolingues ne parvenaient plus à différencier les deux langues alors que les enfants bilingues en étaient toujours capables. Sebastián-Gallés, Albareda-Castellot, Weikum et Werker (2012) ont poussé plus loin cette étude. Tout en utilisant les mêmes stimuli, ils ont testé des bébés de huit mois monolingues (espagnol ou catalan) et bilingues (espagnol-catalan) n'ayant jamais entendu l'anglais ou le français. Les résultats montrent que les monolingues sont incapables de discriminer visuellement deux langues qui ne leur sont pas familières alors que les bilingues en sont capables. Il semblerait donc que les monolingues et les bilingues utilisent des processus différents pour décoder les informations visuelles et que les enfants bilingues soient plus « sensibles » à la cohérence des gestes articulatoires utilisés par une langue donnée. De plus il semblerait que chez les enfants bilingues, le processus d'affinage perceptif (pour les informations visuelles non natives) se met en place plus tardivement. Cependant, ces études concernent la discrimination des langues entre elles et pas la discrimination de phonèmes *per se*.

Chez l'adulte et dans le cadre de la discrimination de phonèmes, Navarra et Soto-Faraco (2007) montrent que des bilingues catalan-espagnols dominant en espagnol échouent à discriminer le contraste catalan /e/-/ɛ/ (qui n'existe pas en espagnol) lors d'une présentation auditive seule. Cependant, lorsque les participants pouvaient s'appuyer sur les mouvements oro-faciaux du locuteur (i.e., en présentation audiovisuelle) ils parvenaient à distinguer les deux phonèmes. Toutefois, deux points sont à souligner ici. D'une part, les phonèmes testés était connus des populations puisque le contraste existe en catalan, qui est une langue connue par les participants. D'autre part, l'absence d'un groupe contrôle constitué de monolingues ne permet pas de savoir si l'avantage audiovisuel est plus important pour une population

familiarisée à plusieurs codes labiaux. Est-ce que les individus bilingues de naissance traitent l'information oro-faciale, notamment lors de l'articulation de phonèmes non natifs, de la même façon que le font les monolingues ? L'étude présentée ici examine l'impact de l'expérience linguistique précoce sur le traitement des informations visuelles dans le cadre de la perception audiovisuelle de phonème non natifs.

4.1.3 OBJECTIFS

Le premier groupe de notre étude est donc composé de bilingues précoces catalan-espagnol ainsi que de bilingues parlant le français et une autre langue. Les participants monolingues étaient quand à eux de langue maternelle française. Dans cette étude, les participants devaient discriminer le contraste bengali /t/-/ʈ/ (constitué des consonnes plosives dentale et rétroflexe). Spécifiquement, le phonème /t/ existe en espagnol et en français ce qui n'est pas le cas de la consonne rétroflexe, qui n'existe dans aucune des langues parlées par les participants. Les enregistrements furent présentés en modalité auditive afin d'examiner les capacités des deux groupes à discriminer auditivement le phonème natif et non natif. Le contraste bengali a également été présenté en modalité audiovisuelle afin de voir si l'information visuelle fournie par les mouvements oro-faciaux contribuait à limiter le phénomène de surdité phonologique. Puisque les monolingues et les bilingues passent par un processus d'affinage perceptif différent pour le décodage de la parole visuelle, il est probable que les processus de traitement de l'information visuelle diffèrent lorsqu'ils sont adultes.

4.2 MATERIEL ET METHODE

4.2.1 PARTICIPANTS

Les informations sur le niveau de maîtrise linguistique des participants ont été collectées avec une version modifiée du questionnaire « *Language Experience and Proficiency Questionnaire* » (LEAP-Q) (Marian, Blumenfeld & Kaushanskaya, 2007 ; Annexe 1). Ce questionnaire édité en plusieurs langues a pour objet la caractérisation du contact de l'individu avec une langue donnée (Marian et al., 2007). Quarante-sept bilingues ont participé à l'expérience. Même si nous n'adressons pas directement la question du bilinguisme précoce et

tardif, nous avons sélectionné des participants ayant été exposés de manière précoce aux deux langues (au plus tard à l'école maternelle) et dont la pratique des deux langues était courante, assurant un niveau de maîtrise « natif ». Vingt-quatre d'entre eux étaient des bilingues catalan-espagnol (quatre hommes et 20 femmes ; *M* = 20 ans ; *SD* = 2 ans). L'âge d'acquisition moyen de l'espagnol et du catalan était de 12 mois. Dix ont appris l'espagnol et le catalan chez eux. Sept ont toujours été exposés à l'espagnol mais vivent à Barcelone et sept ont toujours été exposés au catalan et ont appris l'espagnol à la crèche ou à l'école maternelle. Ils étaient tous étudiants à l'université de Pompeu Fabra (Barcelone, Espagne). Vingt-trois bilingues de langues différentes (huit hommes et 15 femmes ; *M* = 19.6 ans ; *SD* = 3 ans) participaient également à l'étude. L'âge d'acquisition moyen des deux langues était de 14 mois. Ils parlaient tous le français et une autre langue : anglais (4 participants), allemand (4 participants), italien (4 participants), espagnol (3 participants), malgache (2 participants), portugais (2 participants), arabe (2 participants) et polonais (2 participants). Ils ont tous été exposés aux deux langues depuis la naissance. Les parents résidaient à Grenoble mais parlaient leur langue maternelle de façon quotidienne à la maison. Ils étaient étudiants à l'Université de Grenoble ou à la Cité Scolaire Internationale de Grenoble. Le groupe de monolingues était composé de 47 francophones (huit hommes et 39 femme; *M* = 22 ans ; *SD* = 1.7 ans). Ils ont tous appris l'anglais comme deuxième langue au collège et au lycée mais leur maîtrise de la langue était faible. Aucun avait effectué de voyage à l'étranger de plus de un mois. Ils étaient tous étudiants à l'Université de Grenoble et on reçu des crédits académiques pour leur participation.

4.2.2 MATERIEL

Les stimuli étaient issus d'un enregistrement de 20 exemplaires de /ta/ et /ʈa/, deux syllabes du Bengali qui diffèrent en termes de place d'articulation. Alors que la première contient une consonne plosive *dentale* qui existe dans les langues des participants, la seconde est une consonne plosive *rétroflexe* qui n'existe dans aucune des langues parlées par les participants. Ces deux syllabes ne constituent donc un contraste phonologique pour aucun des participants. Les stimuli ont été enregistrés dans une chambre sourde par une locutrice du bengali native du Bangladesh. Le visage complet de la locutrice apparaissait sur un fond bleu (Figure 35). L'enregistrement a été fait avec une caméra tri-CCD SONY DXC-990P et un microphone AKG C1000S. Les fichiers ont par la suite été convertis en format vidéo (format PAL, 25

img/sec) et les stimuli ont été segmentés manuellement avec le logiciel *Dps Reality software*. Chaque stimulus commence et se termine avec la locutrice bouche fermée. Onze exemplaires de chaque séquence ont été sélectionnés sur les 20 enregistrés. La Figure 35 présente les caractéristiques auditives et visuelles des deux types de syllabes utilisées dans l'expérience.

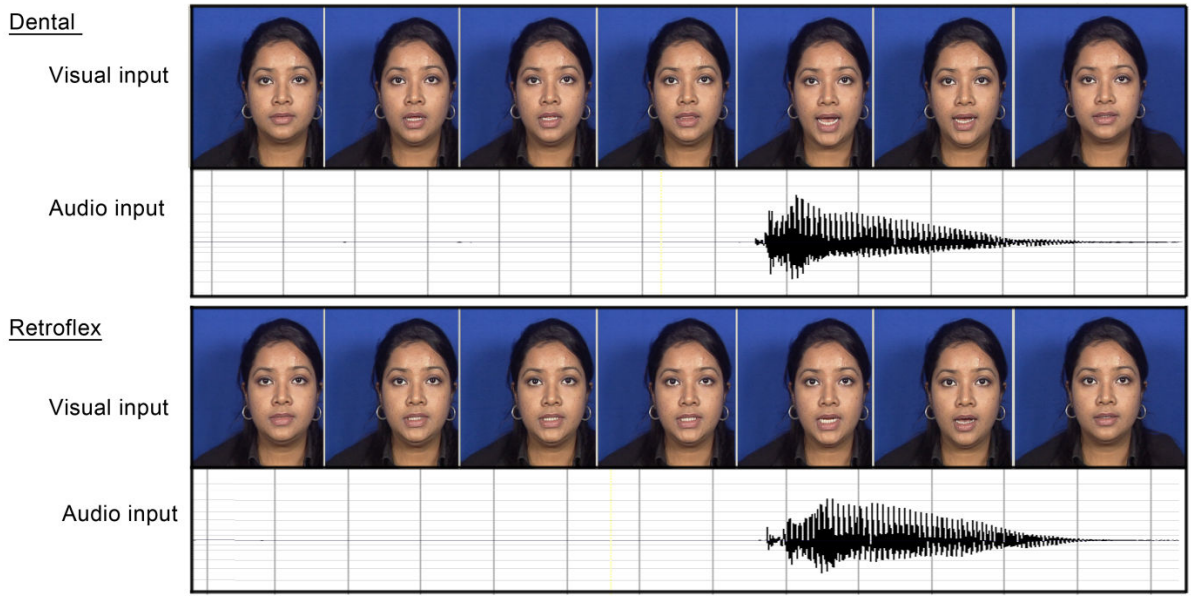

Figure 33. Représentation des caractéristiques auditives et articulatoires durant la production de la syllabe dentale /ta/ et rétroflexe /ʈa/.

4.2.3 PROCEDURE

Un paradigme ABX[39] programmé sur Eprime® software 2.0 (Psychology Software Tools, Inc.) a été utilisé pour cette expérience. Dix exemplaires de /ta/ et 10 de /ʈa/ ont été utilisés comme stimuli X. L'exemplaire restant était utilisé comme stimulus A ou B. Lors de l'expérience, le stimulus auditif était présenté en premier, suivi du B. L'essai se terminait par la présentation du stimulus X. Tous les stimuli étaient présentés durant 1250 ms indépendamment de leur durée de production. Un rappel des touches de réponse était affiché jusqu'à la réponse du participant. Le participants avait pour consigne d'appuyer sur une

[39] Le paradigme ABX consiste à présenter de manière successive une premier stimulus (A), suivi d'un deuxième stimuli (B) qui diffèrent sur une caractéristique (e.g., la place d'articulation dans la cas de présentation auditive). Par la suite, un stimulus X, qui est similaire à A ou B est présenté. La participant, doit, s'il est sensible à la caractéristique manipulée, indiquer si le stimulus X correspond à A ou B. Si le participant n'est pas sensible à la caractéristique manipulée, il répondra au hasard.

touche si la syllabe présentée en X était la même que celle présentée en A, ou sur une autre touche si la syllabe présentée en X était la même que celle présenté en B. Ils devaient répondre aussi précisément et rapidement que possible (même pendant le stimulus). Alors que l'ordre de présentation des stimuli A et B étaient contrebalancés entre les participants, le stimulus X étaient sélectionné aléatoirement parmi les 10 exemplaires de /ta/ (ou /ʈa/) au fur au fur et à mesure des essais.

L'expérience était constituée de deux blocs, un pour chaque modalité de présentation, dont l'ordre était contrebalancé entre les participants. Les participants pouvaient soient entendre (présentation auditive) ou entendre et voir (présentation audiovisuelle) les séquences en Bengali. Durant la présentation auditive, les participants voyaient une photo du visage complet de la locutrice. Lors de la présentation audiovisuelle, ils voyaient une vidéo de la locutrice prononçant les séquences. Entre les deux blocs, un écran apparaissait, signalant aux participants le changement de modalité de présentation des items ; lors de ce changement, il leur était possible de faire une pause avant de poursuivre l'expérience. Chaque sujet voyait 40 (soit 20 x 2) stimuli.

L'expérimentation se déroulait dans un box calme. Les participants étaient placés à 40 cm d'un écran LCD (Dell, 17 pouces) devant lequel était placé un clavier de réponse. Les vidéos étaient présentées en 25 img/s avec une résolution de 720 x 576 pixels. Les présentations auditives étaient faites à 44100 Hz par 2 haut-parleurs SONY SRS-88 placés de chaque côté de l'écran. Les participants étaient instruits que lors de la présentation audiovisuelle, ils devaient écouter et regarder attentivement la vidéo, l'attention portée à l'une ou l'autre modalité pouvant moduler l'intégration audiovisuelle des stimuli (Amano & Sekiyama, 1998 ; Summerfield & McGrath, 1984 ; Tiippana, Sams, & Andersen, 2001). Cependant les participants avaient pour consigne de répondre sur la base de ce qu'ils « percevaient » sans faire référence à l'une ou l'autre des modalités.

Une phase d'entraînement précédait toujours la phase de test et comportait 4 stimuli prononcés par une locutrice vietnamienne (deux items pour chacune des modalités de présentation). La procédure étant strictement identique à la phase de test, hormis le fait que l'expérimentateur était présent, afin de s'assurer que le sujet avait compris la tâche et qu'il s'était familiarisé avec la procédure. La durée totale de l'expérimentation était de 25 minutes ; la phase de consigne et de familiarisation durait 10 minutes et le test 15 minutes environ.

4.3 RESULTATS

Le nombre de réponses correctes et le temps de réponse ont été mesurés. Les temps de réponse inférieurs à 300 ms et supérieurs à 3000ms ont été supprimés de l'analyse (4,41% des données). Les données ont été analysées en utilisant le Modèle Linéaire mixte (Baayen, Davidson & Bates, 2008; Bates, 2005), qui permet de prendre en compte la variabilité des participants et des items. Elles furent réalisées sous R (R Development Core Team, 2009) en utilisant le package lme4 (Bates & Maechler, 2009). L'analyse statistique des réponses correctes et du temps de réaction a été réalisée sur le Groupe (monolingue, bilingue), la Modalité de présentation (auditive, audiovisuelle) et le Type de phonème (natif, non natif).

4.3.1 POURCENTAGE DE REPONSES CORRECTES

L'analyse n'a révélé aucun effet principal. Cependant les interactions de trois facteurs sont significatives. Premièrement, la Modalité de présentation et le Groupe entrent en interaction ($t_{(3757)} = 3.13$, $p < .001$). La figure 36 présente le pourcentage de réponses correctes moyens des monolingues et des bilingues en fonction de la Modalité de présentation (auditive, audiovisuelle).

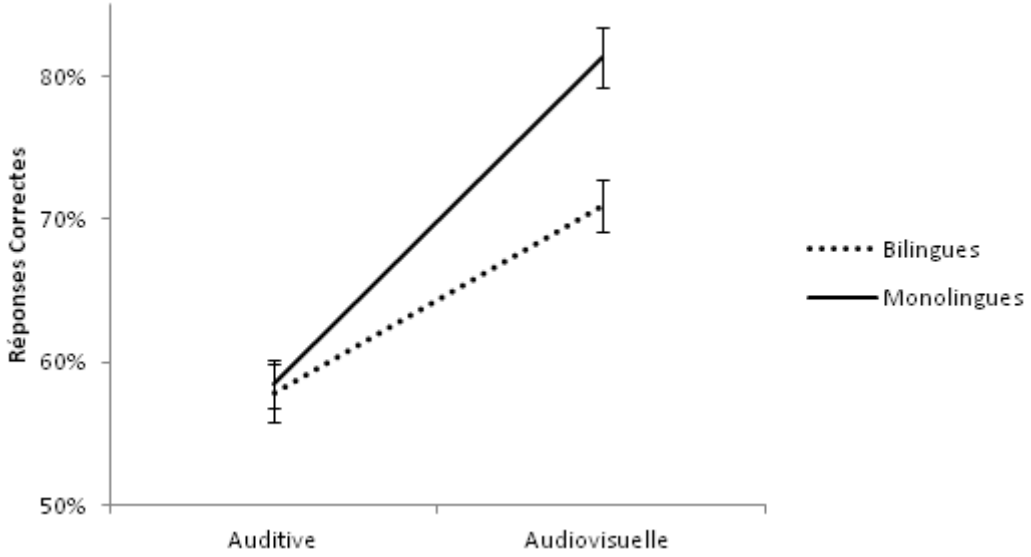

Figure 34. Pourcentage de réponses correctes des monolingues et des bilingues durant la présentation auditive et audiovisuelle du contraste bengali.

Les comparaisons par paires révèlent que les scores des deux groupes augmentent lors de la présentation audiovisuelle par rapport la présentation auditive (monolingues : $M_{auditive}$ =

59% ; $M_{audiovisuelle}$ = 81% , $t_{(1879)}$ = 11.34, p < .001; bilingues : M_{auditive} = 58% ; $M_{audiovisuelle}$ = 71%, $t_{(1879)}$ = 6.07, p < .001). Cependant, le « bénéfice audiovisuel » (score présentation audiovisuelle - score auditive) est plus important pour les monolingues (M = 22%) que pour le bilingues (M= 13%). En présentation audiovisuelle, les monolingues ont plus de réponses correctes que les bilingues ($M_{monolingues}$ = 81% ; $M_{bilingues}$ = 71%, $t_{(1879)}$ = 3.50, p < .001). En revanche, les deux groupes ont des performances équivalentes en présentation auditive ($M_{monolingues}$ = 59% ; $M_{bilingues}$ = 58%, $t_{(1879)}$ = .22, p = .82). De plus les performances des deux groupes diffèrent du hasard (50% de réponses correctes) : pour les monolingues, $t_{(1,46)}$ = 4.65, p < .001 ; pour les bilingues, $t_{(1,46)}$ = 3.71, p < .001.

Deuxièmement, l'interaction entre la Modalité de présentation et le Type de phonème est significative ($t_{(3757)}$ = 3.08, p < .01). La Figure 37 présente le pourcentage de réponses correctes moyen pour les phonèmes natifs et non natif en fonction de la Modalité de présentation (auditive, audiovisuelle).

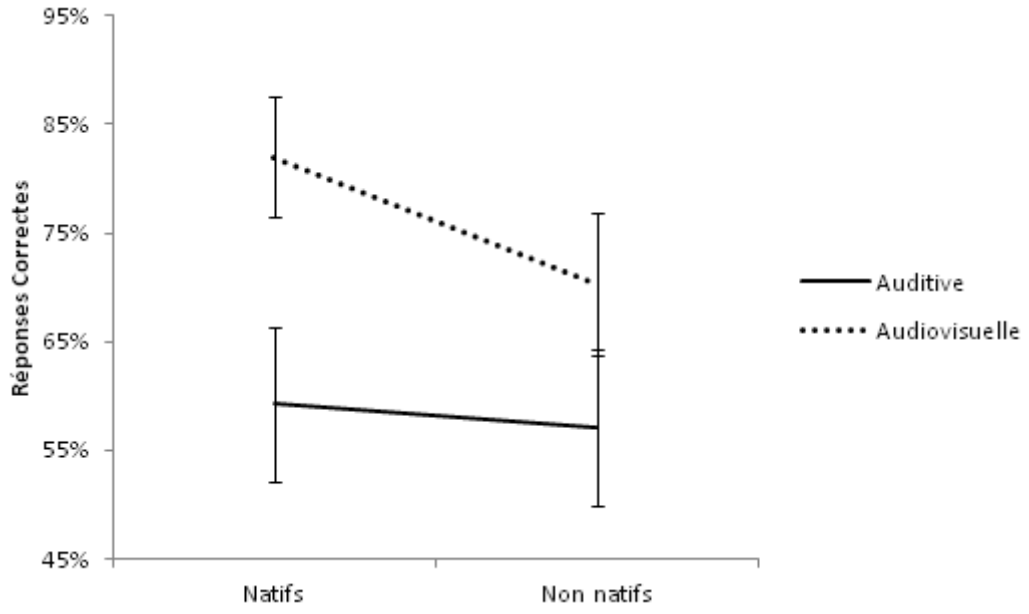

Figure 35. Pourcentage de réponses correctes obtenues lors de la présentation auditive et audiovisuelle des phonèmes natifs et non natifs.

Les comparaisons par paires indiquent des scores plus élevés lors d'une présentation audiovisuelle par rapport à auditive (natifs : $M_{auditive}$=59% ; $M_{audiovisuelle}$ = 82% ,$t_{(1879)}$ = 11.83, p < .001 ; non natif : $M_{auditive}$ = 57% ; $M_{audiovisuelle}$ = 70%, $t_{(1879)}$ = 6.20, p < .001). Cependant, « l'avantage audiovisuel » est plus important pour les phonèmes natifs (M = 22%) que pour les non natifs (M = 13%). Durant la présentation audiovisuelle, les scores obtenues lors de la

présentation des phonèmes natifs sont plus importants que ceux des phonèmes non natifs (M_{natif} = 82% ; $M_{non\ natif}$ = 70%, $t_{(1879)}$ = 4.90, p < .001). Aucune différence entre les phonèmes n'est obtenue en modalité auditive (M_{natif} = 59% ; $M_{non\ natif}$ = 57 %, $t_{(1879)}$ = .84, p = .40).

Troisièmement, l'interaction entre le Groupe et le Type de phonème est significative $t_{(3757)}$ = 2.42, p < .01. La figure 38 présente le pourcentage de réponses correctes moyens pour les phonèmes natifs et non natif en fonction du Groupe (monolingues, bilingues).

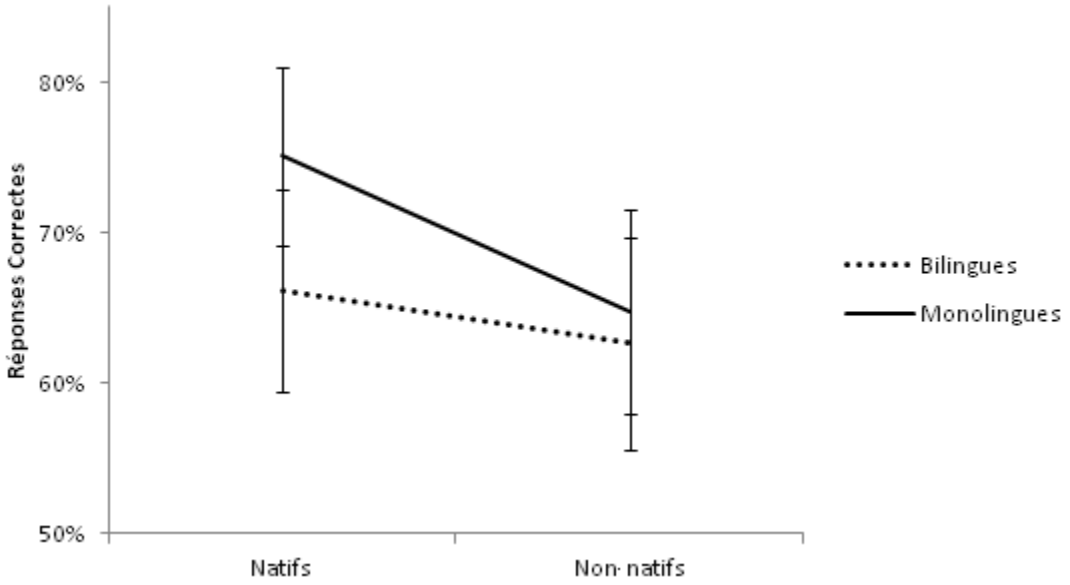

Figure 36. Pourcentage de réponses correctes obtenus lors des présentations auditive et audiovisuelle pour les phonèmes natif et non natifs.

Les comparaisons par paires ne montrent aucune différence significative entre les performances des deux groupes pour les phonèmes non natifs ($M_{monolingues}$ = 65% ; $M_{bilingues}$ = 63%, $t_{(1879)}$ < 1). En revanche, les performances pour les phonèmes natifs sont plus importants pour les monolingues que pour les bilingues ($M_{monolingues}$ = 75% ; $M_{bilingues}$ = 66%, $t_{(1879)}$ = 2.40, p < .01). Pour les monolingues, de meilleurs scores sont obtenus pour les phonèmes natifs que non natifs (M_{natifs} = 75% ; $M_{non\ natif}$ = 65%, $t_{(1879)}$ = 4.34, p < .001) alors que les scores sont similaires pour les bilingues (M_{natifs} = 66% ; $M_{non\ natif}$ = 63%, $t_{(1879)}$ = 1.61, p = .10).

4.3.2 TEMPS DE REPONSES

L'analyse des temps de réactions montre que les monolingues répondent en moyenne plus rapidement que les bilingues ($M_{monolingues}$ =1370ms ; $M_{bilingues}$ = 1488ms, $t_{(3591)}$ = -2.45, p < .01). De plus les temps de réponse fournis lors d'une présentation audiovisuelle sont également plus courts que lors de la présentation auditive ($M_{audiovisuelle}$ =1359ms ; $M_{auditive}$ = 1499ms, $t_{(3591)}$ = -2.45, p < .01). Enfin, les réponses données pour les phonèmes natifs sont plus rapides que pour les phonèmes non natifs (M_{natifs} =1390ms ; $M_{non\ natif}$ = 1464ms, $t_{(3591)}$ = -2.31, p < .05). L'interaction entre la Modalité de présentation et le Groupe est significative ($t_{(3591)}$ = -2.75, p < .001). Aucune des autres interactions ne sont significatives. La Figure 39 présente les Temps de réponse moyens des monolingues et des bilingues en fonction de la Modalité de présentation (auditive, audiovisuelle).

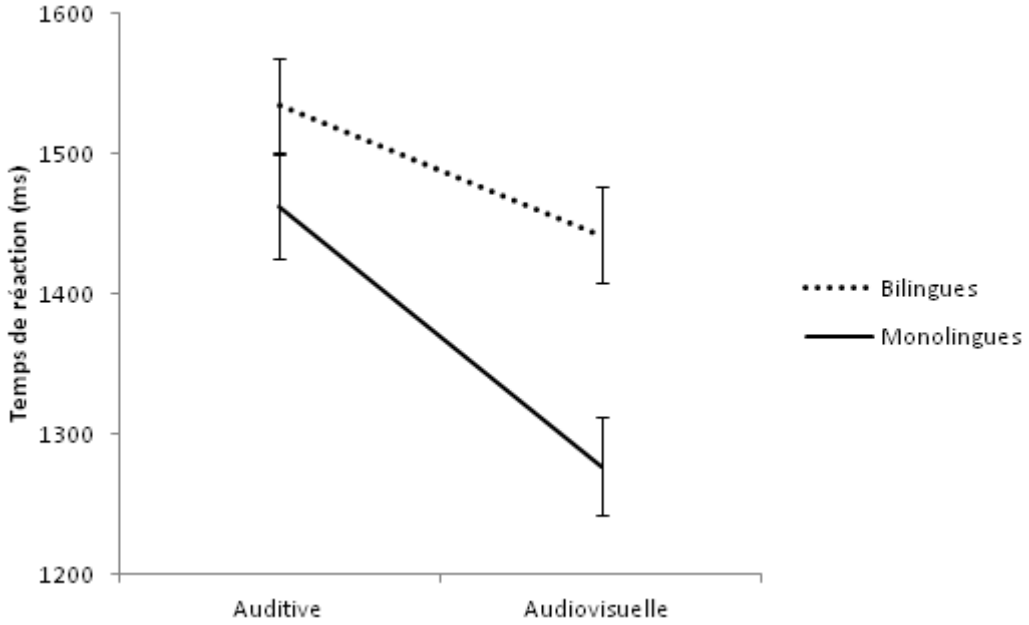

Figure 37. Temps de réaction moyen (ms) pour les réponses correctes pour els monolingues et les bilingues lors de la présentations auditive et audiovisuelle du contraste bengali.

Les comparaisons par paires révèlent que les deux groupes répondent plus rapidement lors d'une présentation audiovisuelle (monolingues : $M_{auditive}$ = 1463 ms ; $M_{audiovisuelle}$ = 1277 ms , t(1822) = -10.22, p < .001; bilingues : $M_{auditive}$ = 1534 ms ; $M_{audiovisuelle}$ = 1442 ms, $t_{(1822)}$ = -4.44, p < .001).Toutefois, le bénéfice temporel induit par la présentation audiovisuelle (Temps de réponse présentation auditive -Temps de Réponse audiovisuelle) est plus important pour les monolingues (186 ms) que pour les bilingues (92 ms). Les monolingues répondent également plus vite que les bilingues lors de la présentation audiovisuelle ($M_{monolingues}$ = 1277 ms ; $M_{bilingues}$ = 1385 ms, $t_{(1791)}$ = -3.20, p < .001) alors qu'aucune différence n'apparaît pour

les stimuli présentés en modalité auditive ($M_{monolingues}$ = 1463 ms ; $M_{bilingues}$ = 1534 ms, $t_{(1791)}$ = -1.39, p = .16).

4.4 DISCUSSION

Le but de cette expérience était de déterminer si les bilingues et le monolingues (qui n'ont pas eu la même expérience linguistique durant l'enfance) tirent le même avantage des informations visuelles sur les mouvements articulatoires du locuteur lorsqu'ils doivent distinguer des phonèmes qui n'existent pas dans leur langue maternelle. Les participants monolingues et bilingues devaient discriminer un contraste consonantique bengali dont les deux phonèmes se distinguaient sur la place d'articulation (i.e., dentale/rétroflexe). Les phonèmes étaient présentés en modalité audiovisuelle durant laquelle une vidéo fournissait à la fois des informations auditives mais également celles concernant les mouvements articulatoires du locuteur en train de produire les séquences. Dans une condition auditive, les mêmes stimuli étaient présentés accompagnés d'une photo de la locutrice. Les résultats indiquent que durant la présentation auditive, les monolingues ainsi que les bilingues ont les mêmes difficultés pour discriminer le phonème rétroflexe du bengali qui n'existe pas dans leurs langue/s. Cependant, lors de la présentation audiovisuelle, les deux groupes tirent partie des informations visuelles fournies par le visage du locuteur pour reconnaître le phonème non natif. Ils peuvent (au moins partiellement) surmonter la surdité phonologique qu'ils subissent lors de la présentation auditive. La présentation audiovisuelle ne fait pas qu'améliorer les performances de discrimination, elle accélère également les traitements. Les résultats ont également fait ressortir des différences dans les traitements visuels effectués par les deux groupes puisque dans la condition audiovisuelle les monolingues ont de meilleurs scores et sont plus rapides que les bilingues. De plus, le « bénéfice audiovisuel » était plus important pour les monolingues que pour les bilingues indiquant que l'exposition précoce à plus d'une langue pourrait affecter la manière dont les traitements visuels sont réalisés, au moins lors de l'identification des phonèmes non natifs.

Les monolingues et les bilingues ont des scores et des temps de réaction similaires lors de la présentation auditive. Cela suggère que lorsque les participants entendent le phonème rétroflexe /ʈ/ qui n'existe pas dans leur répertoire phonologique natif, ils l'assimilent au phonème /t/ qui existe dans leur langue. Ce phénomène d'assimilation entraîne de grandes difficultés lors de la discrimination des phonèmes. Celle-ci survient car les participants

n'arrivent pas à traiter les indices auditifs pertinents qui permettent de différencier les deux phonèmes (Best et al., 2001). Cela est en accord avec de précédentes recherches qui montrent que les bilingues précoces peuvent avoir des performances semblables à celles des monolingues lors de la perception de phonèmes non natifs (Pallier, Bosch, & Sebastián-Gallés, 1997 ; Sebastián-Gallés & Soto-Faraco, 1999). Par exemple, Von Holzen & Mani (2012) montrent que des enfants d'âge préscolaire bilingues français-allemand échouent à discriminer des consonnes salish. De plus, l'étude de Navarra & Soto-Faraco (2007) indique que lors d'une présentation auditive, les bilingues catalan-espagnol dominant espagnol ne peuvent discriminer les phonèmes catalans /e/ et /ɛ/ (seul /e/ existe en espagnol). Cela suggère que la sensibilité particulière qu'ont les bilingues durant leur première année de vie ne s'étend pas nécessairement aux phonèmes non natifs plus tard dans leur vie. Même si plusieurs études sur la perception des enfants montrent une sensibilité aux phonèmes plus accrus chez les bilingues, le développement perceptif continue de changer après la première année de vie (Sundara, Polka, & Genesee, 2006). Nos résultats sont cohérents avec cette idée puisque lors de la présentation auditive, les bilingues ne montrent pas d'avantage particulier lors de la discrimination de phonème par rapport aux monolingues.

La contribution principale de notre étude concerne la composante visuelle du processus de discrimination de phonèmes non natifs. Tous les participants ont réussi à exploiter les indices visuels pour discriminer la rétroflexe /ʈ/ de la dentale /t/ en modalité audiovisuelle. Même si la place rétroflexe n'existe pas dans le répertoire phonologique natif, les différences visuelles entre les deux phonèmes sont suffisamment saillantes pour être utilisées par les participants. Les informations visuelles fournies par l'articulation jouent donc un rôle important pour surmonter, au moins partiellement, les difficultés que les deux groupes subissent lors de la présentation auditive. A notre connaissance, la seule étude sur la perception audiovisuelle des phonèmes par des bilingues a été conduite par Navarra & Soto-Faraco (2007). Comme nous l'avons mentionné, les bilingues catalan-espagnol dominant espagnol testés dans l'étude pouvaient discriminer les phonèmes catalans /e/ et /ɛ/ lors de la présentation audiovisuelle mais pas lors de la présentation auditive. Lors de la présentation audiovisuelle, ils étaient capables de surmonter les difficultés éprouvées lors de la présentation auditive, comme c'est le cas dans la présente étude. Cependant, nous ne savons pas, dans cette étude, si les performances des participants bilingues différaient de celles de monolingues puisque qu'aucun groupe de monolingues n'avait été testé. Notre étude répond donc à cette question.

Les résultats que nous avons obtenus révèlent que les monolingues et les bilingues ne tirent pas le même avantage des informations visuelles sur l'identité des phonèmes. Durant la présentation audiovisuelle, les performances des monolingues sont meilleurs que celles des bilingues. De plus, le « bénéfice audiovisuel » (i.e., l'augmentation du score entre la condition auditive et audiovisuelle) était de 9 % plus élevé chez les monolingues par rapport aux bilingues. Nous avons également pu observer que la présentation des mouvements articulatoires du locuteur accélérait le traitement phonologique pour les deux groupes mais que les bilingues semblent moins sensibles à l'information visuelle. En effet, l'accélération était plus prononcée chez les monolingues que chez les bilingues. La différence de « bénéfice audiovisuel » (i.e., la diminution des temps de réaction) était de 46 ms même si cette différence n'atteint pas la significativité. Cela est consistant avec les résultats de Sebastián-Gallés et al. (2012) avec des enfants de huit mois qui suggèrent que les monolingues et les bilingues utilisent des mécanismes de traitement différents pour décoder la parole visuelle.

Les études sur l'identification/discrimination des phonèmes non natifs chez l'enfant montrent que l'exposition précoce à plusieurs langues retarde l'affinage perceptif et peut entraîner de meilleures performances lors de la perception de phonèmes non natifs (Burns et al., 2003 ; Byers-Heinlein et al., 2010 ; Kuhl et al., 2003). Nos résultats ne confirment pas un avantage pour les bilingues. Etre exposé à plusieurs langues peut retarder l'affinage perceptif mais n'entraîne pas nécessairement un bénéfice quant à l'identification des phonèmes à l'âge adulte. En fait, la facilitation due au bénéfice d'être bilingue pourrait résulter en un *coût cognitif* à l'âge adulte. Par exemple Costa, Caramazza et Sebastian-Galles (2000) fournissent des données de temps de réaction qui indiquent que les bilingues sont systématiquement plus lents que les monolingues lors de tâches de dénomination d'images impliquant des processus d'accès lexical.

Une autre possibilité est que les bilingues tirent moins d'avantages des informations visuelles fournies par les mouvements articulatoires du locuteur car ceux-ci sont mieux équipés pour traiter les informations auditives. Golestani, Molko, Dehaene, LeBihan et Pallier (2007) ont mesuré le gyrus de Heschl chez des participants français qui ont appris un contraste hindi dentale/rétroflexe « lentement » ou « rapidement ». Le gyrus de Heschl est situé dans le cortex auditif et est une des première régions corticales qui reçoit l'information auditive qui vient du système auditif périphérique. Ils observent que les participants qui ont acquis les contrastes rapidement ont un gyrus dont le volume est plus important que les participants qui ont acquis ce contraste lentement. D'après les auteurs, un plus gros gyrus de

Heschl pourrait entraîner une meilleure représentation temporelle des sons. Cela serait très utile pour discriminer les modifications rapides des transitions acoustiques qui sont observées dans de nombreuses consonnes et ainsi, améliorer les capacités de discrimination du contraste dental/rétroflexe. L'étude de Ressel et al. (2012) pointe cette même différence mais entre les monolingues et les bilingues. Ils ont mesuré le volume du gyrus de Heschl chez des monolingues hispanophones et des bilingues catalan-espagnol. Les résultats indiquent que les bilingues ont un gyrus plus gros que les monolingues. Les données volumétriques obtenues pour le gyrus gauche révèlent que la matière grise était présente dans des quantités plus importantes chez les bilingues que chez les monolingues. La corrélation positive entre le volume du gyrus de Heschl et les capacités à percevoir des contrastes non natifs suggèrent que les bilingues auraient de meilleures capacités de discrimination auditive et se reposeraient par conséquent moins sur les informations visuelles. Cependant, même si cette hypothèse semble intéressante, elle n'est pas confortée par nos résultats puisque les bilingues obtiennent le même pattern de résultats que les monolingues lors de la présentation auditive.

Le fait que les monolingues sont plus efficaces que les bilingues pour utiliser les informations visuelles fournies par l'articulation visible pourrait révéler un autre « coût du bilinguisme » qui n'aurait rien à voir avec l'identification des phonèmes *per se* mais avec le traitement visuel du visage du locuteur. En effet, pour décoder la parole visuelle lors de communication face-à-face, nous devons traiter le visage du locuteur. Une analyse de la configuration du visage du locuteur est requise pour localiser la bouche par rapport aux autres éléments du visage (i.e., yeux, nez, ...). Nous serons par la suite en mesure d'analyser les mouvements qui transmettent les informations pertinentes sur l'identité des phonèmes. Cela signifie qu'il pourrait y avoir un lien entre les traitements des visages et le traitement de la parole visuelle. Si c'est le cas, le pattern d'affinage perceptif observé chez les bilingues pour la parole visuelle (Sebastián-Gallés et al., 2012) résulterait-il, ou serait-il relié à l'affinage perceptif subi pour le traitement des visages ? D'un point de vue développemental, le traitement des visages et l'identification des phonèmes sont tous deux importants pour la communication entre le bébé et la personne qui en prend soin. Le visage peut être considéré comme un canal important même avant le début de la mise en place du langage gestuel ou oral (Pascalis et al., 2014). Cette idée d'un lien entre la lecture labiale et le traitement des visages n'est pas nouvelle. En 1986, le modèle de reconnaissance des visages de Bruce et Young incluait déjà un module optionnel d'analyse de la parole faciale qui catégorisaient les mouvements oro-faciaux.

De plus, quelques études suggèrent qu'un environnement bilingue pourrait entraîner des changements dans l'organisation cérébrale et affecter les tâches de perception des visages et de localisation spatiales qui sont liées à une asymétrie hémisphérique (Sewell & Panou, 1983). Plus récemment, Hausmann, Durmusoglu, Yazgan et Güntürkün (2004) ont étudié les différences de spécialisation hémisphériques entre des monolingues allemands et des bilingues turque/allemand durant une tâche linguistique et une tâche de discrimination de visages. Les résultats indiquent que les bilingues n'ont pas le même avantage du champ visuel gauche que les monolingues durant la discrimination de visages. Les temps de réaction des bilingues étaient plus importants que ceux des monolingues quand le visage était présenté dans le champ visuel gauche, reflétant une différence dans l'organisation cérébrale du traitement des visages entre les deux populations. Cette différence temporelle est consistante avec nos données. Nous avons en effet observé que les bilingues étaient plus lents que les monolingues dans la condition audiovisuelle. A partir de cette observation, et si les monolingues et les bilingues traitent les visages différemment, cela peut impacter la façon dont ils utilisent les informations visuelles fournies par les mouvements oro-faciaux du locuteur.

4.5 CONCLUSION GENERALE

Cette étude a permis de mettre en évidence que l'expérience linguistique a un impact sur la façon dont nous traitons la parole visuelle. Les monolingues sont plus performants et rapides que les bilingues pour exploiter les mouvements articulatoires. Cela leur confère un avantage lors de la discrimination de phonèmes non natifs. D'autres recherches seront nécessaires pour savoir si les performances plus faibles et les temps de réaction plus importants des bilingues lors de la présentation audiovisuelle sont dus à des différences lors du traitement du visage.

4.6 RESUME

But de l'étude : Déterminer l'impact de la maîtrise de plusieurs langues sur la capacités à discriminer auditivement et audiovisuellement des phonèmes inconnus.

Populations : 47 bilingues précoces (catalan-espagnol ou français et une autre langue) et 47 monolingues francophones.

Protocole : Un paradigme ABX à été utilisé. Les stimuli A et B étaient des syllabes du Bengali qui varient sur la place d'articulation : /ta/ ou /ʈa/. Les stimuli étaient présentés en modalité auditive et audiovisuelle.

Résultats : L'analyse des réponses correctes a mis en évidence des résultats similaires entre les deux groupes lors de la présentation auditive (tous Types de stimuli confondus), mais un avantage pour les monolingues lors de la présentation audiovisuelle. Le bénéfice audiovisuel est plus important pour les monolingues. La présentation audiovisuelle (tous groupes confondus) permet d'améliorer les performances, mais cet avantage est plus important pour les syllabes contenants le phonème natif. Concernant les temps de réponse, un avantage audiovisuel global a été observé. Les monolingues répondent plus rapidement que les bilingues.

Conclusion : Nous avons mis en évidence des différences entre les deux groupes dans les traitements effectués lors de la présentation audiovisuelle de phonèmes. Un désavantage a été observé aussi bien au niveau de la précision que des temps de réaction pour les bilingues. Ces différences semblent induites par la présentation simultanée des informations visuelle et auditive puisque les deux groupes ne diffèrent pas lors de la présentation auditive, ni en terme de précision ou de temps de réponse. Il semble possible que les deux populations mettent en place des traitements différents pour décoder les mouvements oro-faciaux du locuteur.

CHAPITRE 6

ETUDE 1

EVOLUTION DU DECOURS TEMPOREL DE L'IDENTIFICATION DE PHONEMES NATIFS EN FONCTION DE L'INFORMATION AUDITIVE ET VISUELLE FOURNIE PAR LE LOCUTEUR

5.1 INTRODUCTION

De nombreuses études ont permis de mettre en évidence que l'accès à la double information de parole (i.e., signal acoustique et mouvements articulatoires) augmente l'intelligibilité de la parole dans le bruit (Kim & Davis, 2004) et même lorsque les conditions d'écoute sont adéquates (Reisberg, et al., 1987). Cependant cela ne signifie pas que les deux informations contribuent de la même manière à la reconnaissance de mots ou de phonèmes. D'une part, les informations visuelles, même si elles donnent des indices qui sont *redondants* à ceux fournis par le canal auditif, donnent également accès à des indices *complémentaires* au signal acoustique de parole. Par exemple, alors que le mode d'articulation est un paramètre fortement discriminant au niveau auditif, la place d'articulation est quant à elle très visible (Walden et al., 1975).

D'autre part, cette contribution différentielle des informations visuelles et auditives s'observe car les informations visuelles précèdent, dans plusieurs situations, les informations acoustiques. Des études ont montré que celles-ci étaient présentes et utilisées jusqu'à 40 ms avant l'information auditive pour des consonnes dans des séquences aCa (Schwartz & Savariaux, 2013) et jusqu'à 300 ms pour les séquences à consonne initiale (Chandrasekaran et al., 2009 ; Schwartz & Savariaux, 2013), notamment dans le cas de consonnes biliabiales comme /p/. Pour les voyelles, 160 ms d'avance du signal visuel sur le signal acoustique ont été observée (Cathiard, Lallouache, Mohamadi, & Abry, 1995) même si les informations auditives (notamment avec la voyelle arrondie /y/) peuvent parfois permettre une identification plus précoce que pour la modalité audiovisuelle (Troille et al., 2010) (cf. Chapitre 2.2.1 « Décours temporel du signal de parole audiovisuelle » pour plus de détails).

Enfin les mouvements articulatoires varient en terme de saillance et ne sont pas tous visibles de la même façon. Par exemple la vibration des cordes vocales qui permet de distinguer /p/ de /b/ n'est que rarement visible (Yehia, Rubin, & Vatikiotis-Bateson, 1998). Plus le phonème recrute d'articulateurs visibles, plus l'anticipation articulatoire permettra de prédire le signal auditif qui suivra (van Wassenhove et al., 2005 ; Arnal et al., 2009), notamment via l'accès aux indices de formes (i.e., quelle est l'identité de la consonne qui va suivre) et aux indices temporels (i.e., quand l'information acoustique va débuter). De la saillance visuelle dépendra l'informativité du canal visuel ce qui modulera donc l'utilisation préférentielle d'une où l'autre des informations. En effet, comme l'atteste l'étude de Munhall & Tokura (1998), il semblerait que la valeur informative de chacune des modalités varie en fonction du temps, puisque les indices sur l'identité du phonème ne sont pas présents au même

moment dans les deux modalités. Par exemple, alors qu'il est nécessaire d'attendre le *burst* acoustique (qui est relâché après une période de silence) pour assurer la reconnaissance auditive de /p/, au niveau visuel la fermeture labiale est un indice suffisant pour identifier /p/ parmi des /d/ par exemple.

Toutes ces spécificités permettent à l'avantage audiovisuel d'émerger, celui-ci étant modulé dans le temps et dépend donc de la distribution différentielle des informations fournies par les deux modalités (Jesse & Massaro, 2010). Le but de ce chapitre est d'observer, pour les sons natifs, quel est le décours temporel fin du processus d'identification de phonèmes consonantiques et vocaliques en fonction de la quantité et du type d'information acoustique et/ou visuelle fourni par le locuteur.

5.1.1 OBJECTIFS

Cette expérience a pour but d'examiner l'évolution de l'identification de phonèmes au travers du temps pour un panel de trois consonnes du français qui varient en terme de saillance perceptive. Celle-ci a également pour but de valider un protocole expérimental (déjà utilisé dans la littérature pour décrire le décours temporel de l'utilisation de différente modalité) qui aura à cette occasion été modifié afin de fournir des mesures complémentaires (notamment des temps de réaction). Au travers d'un paradigme de *gating on-line*, nous souhaitons observer la contribution des informations visuelles au regard des informations auditives, afin de déterminer le profil temporel de l'influence de chacune des modalités sur les capacités de détection, notamment en évaluant l'informativité des deux canaux. En effet, nous savons que les informations auditives et visuelles ne sont pas forcément disponibles au même moment, les informations visuelles étant souvent accessibles avant les informations auditives en perception et en production (Abry et al., 1996 ; Chandrasekaran et al., 2009 ; Munhall & Tohkura, 1998 ; Schwartz & Savariaux, 2014 ; Smeele, 1994). L'intégration audiovisuelle est donc dans ce sens un processus temporel modulé par la saillance perceptive des deux types d'informations disponibles, et relève d'une course temporelle dynamique. Il a donc été avancé que l'information visuelle serait exploitée dès qu'elle est disponible et que l'identification des phonèmes serait facilitée par des mouvements articulatoires plus saillants, avant même que l'information auditive ne soit disponible.

D'après Grant et collaborateurs (Grant, Walden, & Seitz, 1998 ; Grant & Walden, 1996), lorsque les modalités auditive et visuelle fournissent les mêmes indices liés à l'identité du phonème, un faible avantage est observé en présentation audiovisuelle. *A contrario*, l'avantage devrait être important quand les modalités fournissent des informations différentes/complémentaires. Une anticipation sera permise si les indices visuels précèdent le flux acoustique (Chandrasekaran et al., 2009) et s'ils sont assez saillants pour pouvoir être utilisés (Smeele, 1994). Le gain audiovisuel sera donc modulé par la saillance visuelle des phonèmes à identifier (van Wassenhove et al., 2005; Arnal et al., 2009). Les consonnes articulées à l'avant du conduit vocal devraient bénéficier d'un large avantage audiovisuel (Smeele, 1994) par rapport à la condition auditive, mais également d'un fort impact de la modalité visuelle, surtout au début de la séquence (Jesse & Massaro, 2010) alors que les places plus en arrière devraient amener une augmentation graduelle du taux d'identification. Nous faisons l'hypothèse qu'un gain audiovisuel sera donc observé quand les deux modalités seront complémentaires, celui-ci se manifestant par une identification plus précoce lors de la présentation audiovisuelle ainsi que par des performances de détection plus importantes lors de la présentation bimodale. Cependant, le bénéfice audiovisuel devrait être moins important, voire inexistant, quand l'une ou l'autre des modalités contient assez d'informations pour permettre l'identification.

Pour mettre en évidence un avantage lors de la présentation audiovisuelle, nous avons utilisé un protocole expérimental de *gating* ou « dévoilement progressif » (Grosjean, 1980). Il s'agit de présenter le signal auditif/audiovisuel/visuel en segments dont la durée est croissante afin de dévoiler de plus en plus de signal. Ce paradigme permet de contrôler la disponibilité des informations à chaque étape durant la présentation. Il est particulièrement intéressant d'exploiter cette spécificité dans le cadre de la parole audiovisuelle puisque les informations auditive et visuelle ne se dévoilent pas au même moment. Cela permet donc de savoir exactement quelle information est perçue par le participant et d'observer si cette information participe ou non au processus d'identification. Ce paradigme a déjà permis de montrer que l'identification, de phonèmes ou de mots, est plus précoce lors de la présentation audiovisuelle (Smeele, 1994 ; Munhall & Tohkura, 1998 ; Jesse & Massaro, 2010 ; Troille et al., 2010 ; Moradi et al. 2013 ; de la Vaux et Massaro, 2004 ; Cathiard, 2010) surtout pour les phonèmes articulés à l'avant du conduit vocal comme les bilabiales et les labiodentales (Smeele, 1994).

Nos études se détacheront de celles précédemment réalisées car elles utilisent un matériel permettant d'améliorer la résolution temporelle des observations. Les stimuli ont été

élaborés à l'aide d'une caméra rapide (100i/s). Les séquences seront présentées par pas de 10 ms, au lieu des 40 ou 25 ms dans les travaux précédents. Cette meilleure résolution temporelle nous permettra de clarifier les apports de chacune des modalités au travers du temps et d'observer des variations fines, notamment sur les seuils de détection. Afin d'obtenir des informations sur l'évolution des temps de réponse au fur et à mesure du dévoilement d'indices acoustiques et visuels, nous avons également rendu la tâche de *gating* on-line, au lieu d'*off-line* dans les études antérieures, en incitant les participants à répondre pendant la présentation des séquences.

5.1.2 PLOSIVES /p t k/

Nous étudierons dans un premier temps les phonèmes des trois consonnes plosives non voisées /p-t-k/. Elles se placent sur un continuum de saillance visuelle (van Wassenhove et al., 2005). En effet /p/ est une consonne bilabiale, par conséquent très visible. Les consonnes /t/ et /k/ ont une visibilité moindre : /t/ étant une labiodentale articulée plus en arrière du conduit vocal et /k/ une vélaire dont l'articulation est presque invisible. Nous avons utilisé des séquences naturelles de type /aCa/ où C varie en terme de saillance visuelle. Les effets de coarticulation ont été maintenus constants en utilisant toujours le contexte vocalique /a/. Le /a/ a été choisi car il affecte peu l'intelligibilité des phonèmes. En effet, Benoît et al. (1994) ont montré qu'une consonne présentée en condition audiovisuelle était mieux perçue dans un contexte /a/ qu'entourée par d'autres voyelles. De plus, il était important de placer une voyelle à l'attaque de la séquence car la parole est la plupart du temps produite de façon continue et la coarticulation qui précède la consonne contient déjà des indices sur l'identité de la consonne suivante (Smits et al., 2003). En effet, les transitions formantiques liées aux contraintes du conduit vocal imposées par la production d'une plosive apparaissent généralement dans les 40 dernières millisecondes de la production de la voyelle précédente mais également dans la voyelle qui succède la consonne. Ces transitions sont spécifiques à chaque consonne comme le montre la Figure 40 et permettent de distinguer certaines consonnes comme /r/-/l/ ou /d/-/g/ (Harris, 1957 ; Mann, 1986).

.

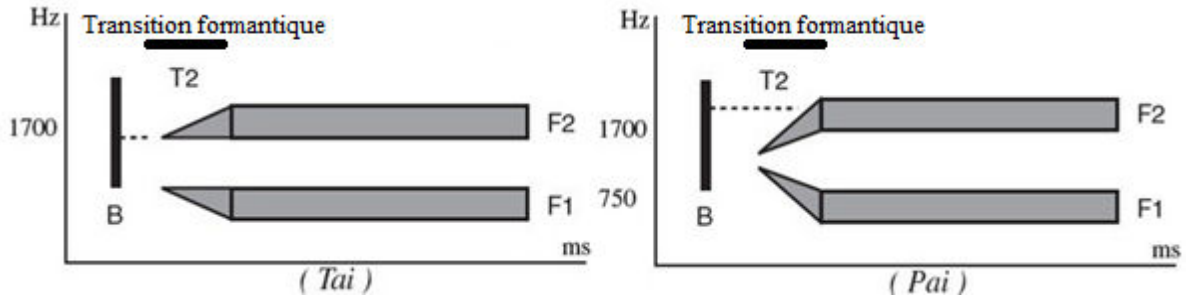

Figure 38. Sonagrammes schématiques des modulations de F1 et F2 pour les consonnes /t/ (transition droite) et /p/(transition avec pente). (Tiré de Virole, 2006)

Ces informations peuvent donc également être exploitées pour identifier un phonème ou au moins participer à l'accumulation d'indices nécessaires à l'identification.

Ces caractéristiques visuelles sont également complétées par des indices acoustiques propres, comme les transitions formantiques. Avec les changements de forme et de taille du conduit vocal lors de la réalisation d'une consonne, la résonance du conduit vocal change. Ces changements induisent les modifications des formants qui contribuent à l'identification de la place d'articulation (Delattre, 1955). Par exemple, /pa/ se caractérise par une augmentation de F2 et F3, /ta/ par une légère diminution de F2 et F3 (Smeele, 1994). Ces indices sont susceptibles d'être utilisés de façon précoce. Ils seront accompagnés d'informations visuelles spécifiques en fonction de la consonne concernée, comme un fermeture labiale plus complète pour /p/ que /t/ par exemple.

5.2 MATERIEL ET METHODE

5.2.1 PARTICIPANTS

Trente-cinq participants (25 femmes et 10 hommes) âgés de18 à 23 ans (*M* = 20 ans) ont pris part à cette expérience. Tous étaient de langue maternelle française et ne présentaient aucun déficit, ni auditif ni visuel ou bien si les sujets avaient un déficit, il était corrigé. Ces participants étaient tous étudiants à la faculté de psychologie de Grenoble et recevaient des crédits académiques en échange de leur participation.

5.2.2 MATERIEL

Les stimuli étaient des enregistrements vidéos de parole naturelle produits par une locutrice de langue maternelle française. Elle a été enregistrée de face avec un fond bleu dans une chambre sourde à l'aide de la caméra rapide S-PRI (cadence 100i/s). Les stimuli consistaient en des séquences CVC qui variaient sur la consonne et la seconde voyelle. Ces stimuli étaient tirés d'un large panel de séquences contenant neuf consonnes (/p-t-k-f-s-ʃ-l-R-h/) et six voyelles du français (/a-e-i-o-u-y/). Les lèvres de la locutrice ont été peintes en bleu afin de pouvoir mesurer les paramètres sur les mouvements labiaux durant la production des séquences via le logiciel TACLE (Traitement Automatique des Contours des Lèvres). Ces données sur la réalisation articulatoire seront exploitées pour la sélection des séquences, ainsi que pour l'interprétation des résultats.

Durant l'enregistrement, la caméra nous fournissait des bips de synchronisation au début et à la fin de chaque séquence ce qui permettait de couper le signal auditif en conservant la synchronisation. Le signal acoustique était enregistré à l'aide d'un appareil d'enregistrement portable (PMD Marantz 670). Quatre à cinq exemplaires de chaque séquence ont été enregistrés, puis segmentés. Nous étudierons les trois consonnes plosives /p t k/ car celles-ci offrent un continuum de saillance et bénéficient d'un indice claire quant à leur commencement (i.e., le *burst*). De plus elles ont été largement étudiées dans la littérature. Après la segmentation, la sélection du matériel a été faite en fonction (1) de la réalisation articulatoire des séquences, ainsi que (2) de leur cohérence temporelle. Les séquences contenant des clignements oculaires ou des bruits de bouche ont été écartées. La qualité de la réalisation articulatoire a été déterminée par analyse de la courbe articulatoire (i.e., paramètre S d'aire aux lèvres, cf. encart supérieur des Figures 41, 42 et 43). Les exemplaires considérés comme moins prototypiques en terme de réalisation ont été rejetés pour la suite des analyses. La cohérence temporelle était quant à elle déterminée par deux indices : le début de fermeture labiale (moment à partir duquel la fermeture labiale est effectuée à hauteur de 10%) et le *burst* acoustique (déterminé visuellement par inspection du signal en utilisant le logiciel Pratt, version 5.1.42 ; Boersma & Weenink, 2010). Seuls les exemplaires dont la durée entre le début de fermeture labiale et le *burst* acoustique étaient équivalentes ont été conservés. Cette durée était en moyenne de 308 ms (SD = 10.14 ms) pour les trois consonnes. Le début de la fermeture était présentée au *gate* 110. Cette mesure, ainsi que d'autre contrôles étaient nécessaires afin que les participants ne basent pas leur jugement sur des indices inappropriés. Ainsi, la durée entre la fin acoustique de la voyelle (*gate* 170) et le *burst* acoustique ont

également été contrôlées. Cette durée était en moyenne de 203 ms (SD = 7.21 ms) pour chaque séquence. Un exemplaire de chaque séquence d'intérêt (i.e., /apa/, /ata/ et /aka/) a été sélectionné (Figures 41, 42 et 43). Les séquences ont par la suite été segmentées selon les contraintes du *gating*. Le premier *gate* joué était de 50 ms. Celui-ci était plus long car il était constitué d'un *fade-in* visuel de 50 ms qui avait pour but de limiter les phénomènes de reconnaissance précoce liés à des caractéristiques visuelles (tels que l'ouverture de la bouche de la locutrice qui pouvait varier au début des séquences). La séquence débutait 370 ms avant le *burst* afin que le début de fermeture soit visible (ce point étant considéré comme le commencement potentiel des informations visuelles sur l'identité de la consonne). Le *burst* était le point acoustique sur lequel les séquences étaient synchronisées, puisqu'il devait assurer la reconnaissance auditive des stimuli. Les *gates* situés avant et après ce point étaient créés en supprimant ou ajoutant 10 ms de signal. La fin de la séquence se situait à 120 ms après le *burst* acoustique, ce qui permettait d'entendre la seconde voyelle, assurant ainsi l'identification de la consonne. Le nombre de *gates* ainsi que le *gate* contenant le *burst* acoustique étaient ainsi le même pour tous les exemplaires (i.e., 470 ms). Notons que dans la suite de la section, nous ferons référence au *gate* par la durée de signal qu'il contient et non par sa numérotation absolue, par exemple, *gate* 370 pour indiquer le *gate* contenant le *burst* et pas *gate* n°320. Les Figures suivantes présentent le détail des stimuli (Figure 41, 42 et 43).

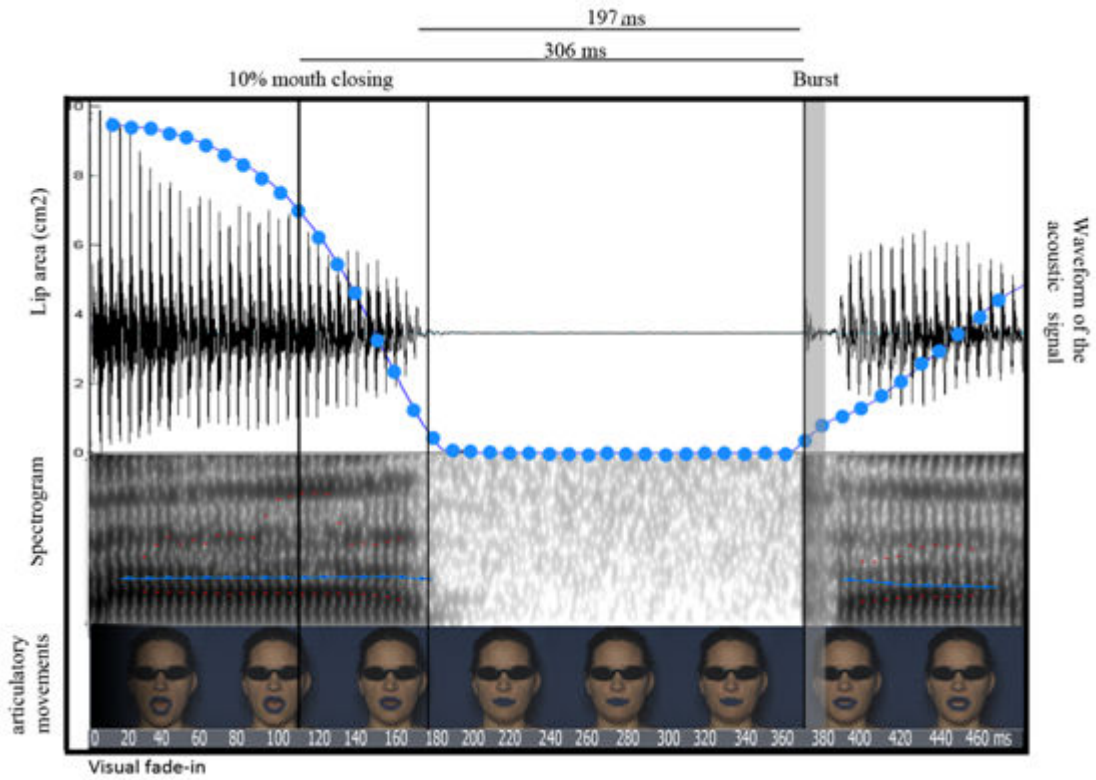

Figure 39. Représentation du décours temporel des informations acoustique et articulatoire pour la séquence /apa/. L'encart supérieur montre l'aire aux lèvres (S ; points bleu) et le signal acoustique (en noir) qui ont servis de base pour la sélection et la découpe des séquences. L'encart inférieur contient le spectrogramme du signal acoustique. Le *burst* (*gate* 370) est le point sur lequel les signaux ont été alignés. Le point où 10% de fermeture labiale (*gate* 110) est atteint à été considéré comme le moment où l'information visuelle sur l'identité de la consonne est disponible. Le *gate* 170 correspond quant à lui à la fin acoustique de la première voyelle, qui coïncide ici avec la fermeture. Les durées entre ces indices sont indiquées dans la partie supérieure de la Figure.

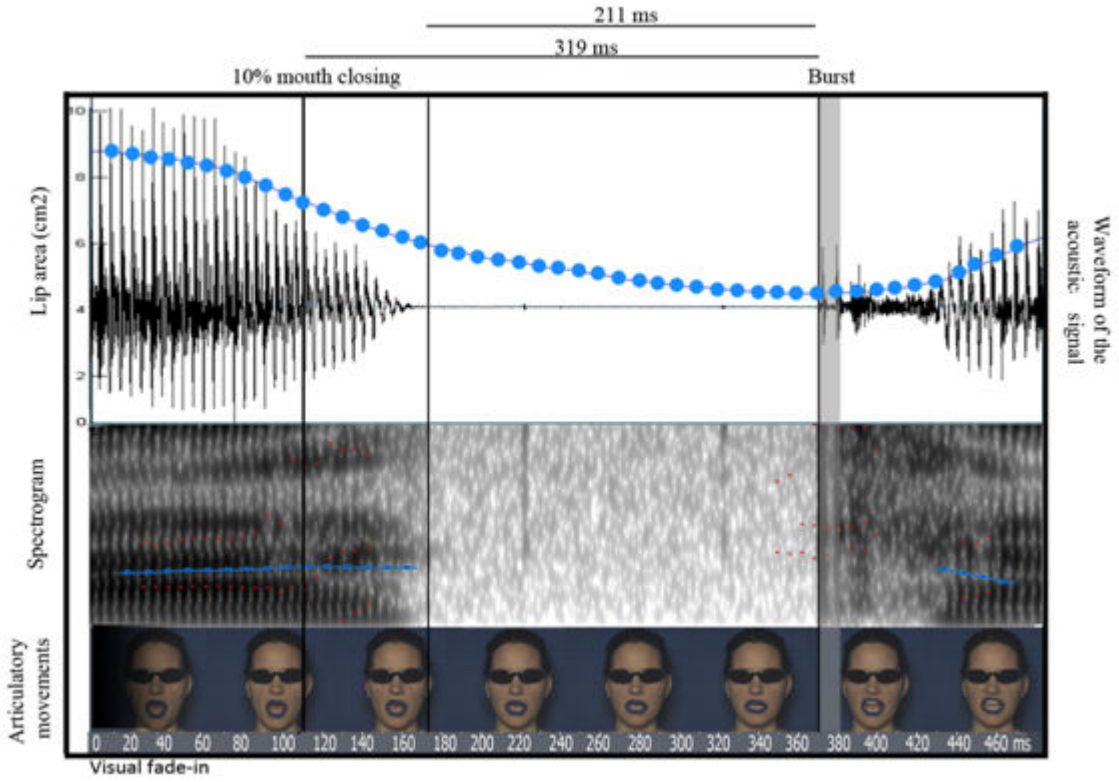

Figure 40. Représentation du décours temporel des informations acoustique et articulatoire pour la séquence /ata/. L'encart supérieur montre l'aire aux lèvres (S ; points bleu) et le signal acoustique (en noir) qui ont servis de base pour la sélection et la découpe des séquences. L'encart inférieur contient le spectrogramme du signal acoustique. Le *burst* (*gate* 370) est le point sur lequel les signaux ont été alignés. Le point où 10% de fermeture labiale (*gate* 110) est atteint à été considéré comme le moment où l'information visuelle sur l'identité de la consonne est disponible. Le *gate* 170 correspond quant à lui à la fin acoustique de la première voyelle. Les durées entre ces indices sont indiquées dans la partie supérieure de la Figure.

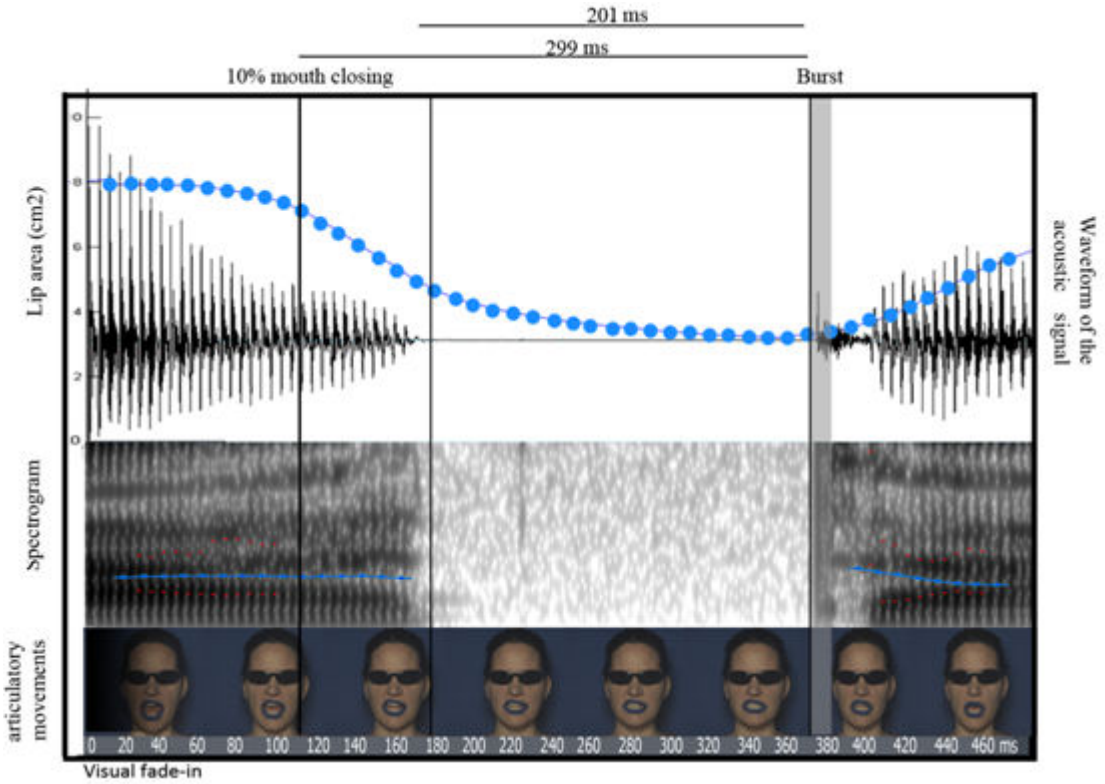

Figure 41. Représentation du décours temporel des informations acoustique et articulatoire pour la séquence /aka/. L'encart supérieur montre l'aire aux lèvres (S ; points bleu) et le signal acoustique (en noir) qui ont servis de base pour la sélection et la découpe des séquences. L'encart inférieur contient le spectrogramme du signal acoustique. Le *burst* (*gate* 370) est le point sur lequel les signaux ont été alignés. Le point où 10% de fermeture labiale (*gate* 110) est atteint à été considéré comme le moment où l'information visuelle sur l'identité de la consonne est disponible. Le *gate* 170 correspond quant à lui à la fin acoustique de la première voyelle. Les durées entre ces indices sont indiquées dans la partie supérieure de la Figure.

5.2.3 PROCEDURE

Contrairement à une tâche de *gating* classique, qui est une tâche dite *off-line* car les participants doivent donner leur réponse après la présentation complète du stimulus, nous avons modifié la tâche afin de la rendre *on-line*. Avec cette variante du paradigme de *gating*, les participants n'attendent pas la fin du stimulus pour répondre mais doivent fournir leur réponse au moment même où ils identifient le phonème cible, ce qui nous permet, à travers la mesure des temps de réaction, d'obtenir des informations sur les processus en temps réel (i.e., pendant que le processus a lieu). Avec cette modification, il n'était pas possible de présenter tous les stimuli dans une liste contenant les trois consonnes /p/-/t/-/k/ et de faire une tâche d'identification comme c'est généralement le cas dans les protocoles de *gating*. Nous avons

donc mis en place une tâche de détection de phonèmes de type *go/no go*[40] (Figure 44). Tous les *gates* des trois séquences /apa/-/ata/-/aka/ étaient présentés aléatoirement dans une liste et une seule consonne (soit /p/, soit /t/, soit /k/), annoncée au préalable, était à détecter. Cette liste était donc jouée trois fois pour que les trois consonnes soient à détecter consécutivement.

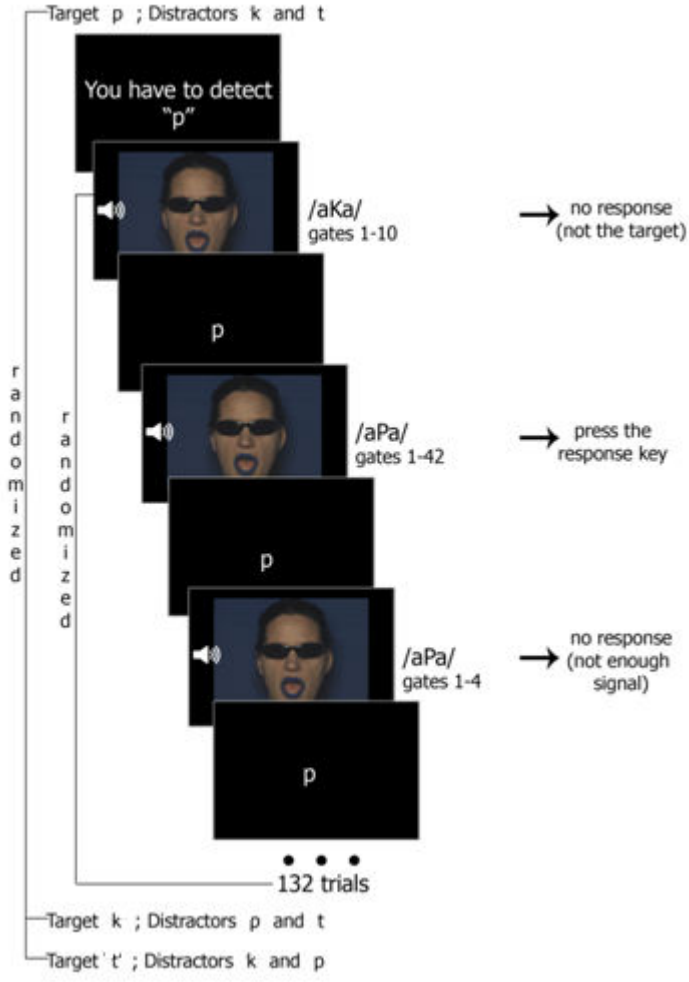

Figure 42. Représentation schématique de la procédure expérimentale pour le bloc audiovisuel.

Le participant est assis face à un ordinateur. Les stimuli visuels et audiovisuels étaient présentés au centre d'un écran de 24" GD254hq-ACER sur un fond noir. Le son était diffusé à un volume confortable via deux enceintes situées de chaque coté du moniteur. Durant la présentation audiovisuelle, les séquences étaient présentées sous forme d'images successives de 800x600pxl à raison d'une image toutes les 10ms. Le logiciel Eprime permettant un chargement en ligne des images et du son qui permettait une synchronisation entre les deux canaux de +/- 1ms. Durant la condition Audio-seule, une croix de fixation apparaissait à l'écran durant la diffusion de la séquence, et en condition visuelle-seule, seule les images étaient présentées.

Les trois modalités de présentation constituaient des blocs différents. L'expérience commençait par la modalité auditive seule ou audiovisuelle, l'ordre de celles-ci étant contrebalancé entre les participants. La condition visuelle était toujours la dernière modalité testée. Pour chaque bloc, les participants étaient exposés à la même liste trois fois. La liste contenait tous les *gates* (du *gate* 50 au *gate* 470) de /apa/, /ata/ ou /aka/ (soit 135 essais par

[40] Le *go/no go* est un procédure dans laquelle le participant doit répondre lorsqu'il perçoit le stimulus cible ("*go*") et ne rien faire s'il ne le perçoit pas ("*no go*").

liste, 405 essais par bloc) qui étaient présentés de manière aléatoire. A chaque fois qu'une liste était jouée, la consonne à détecter changeait. Par exemple durant le bloc audiovisuel, le participant était informé qu'il allait devoir détecter une consonne spécifique (par exemple « vous devez détecter /p/ ») dans des distracteurs (/t/ et /k/). Par la suite, la même liste était jouée une nouvelle fois avec un des distracteurs précédents qui devenait la consonne à détecter (/p/ devient un distracteur et /k/ devient le phonème cible). L'ordre des cibles à l'intérieur des blocs était aléatoire. Une fois les trois listes présentées et le bloc audiovisuel terminé, la modalité de présentation changeait et les trois listes étaient à nouveau présentées aux participants. La tâche des participants était de répondre lorsqu'ils détectaient la consonne cible et de ne rien faire (1) s'ils n'avaient pas un signal assez long pour identifier la consonne ou (2) si la consonne perçue n'était pas la consonne cible. Nous encouragions les participants à ne pas attendre que toute la séquence soit dévoilée pour répondre, et leur indiquions qu'ils pouvaient baser leur réponse sur une intuition.

Avant l'expérience, le participant réalisait un entraînement sur d'autres consonnes que celles de l'expérience (i.e., /asa/, /afa/ and /aʃa/). L'expérience durait environ une heure.

5.2.4 CONDITIONS DE REJETS DE PARTICIPANTS

Neuf participants sur 35 ont été écartés pour les analyses. Afin de limiter les biais par des facteurs autres que ceux que nous désirons observer, nous avons visuellement inspecté les patterns individuels. Ainsi, les participants qui détectaient la consonne (1) à partir des *gates* 50 ou 60, durant au moins 4 *gates* successifs (2) et ce, pour les deux modalités impliquant les même indices (**A**V et **A** / A**V** et **V**). Des détections précoces (avant 60 ms) lors des présentations audiovisuelle et visuelle indiquaient donc l'utilisation d'un indice visuel (e.g., la forme de la bouche de la locutrice) qui aurait été détecté et mémorisé au fur et à mesure des essais. Inversement, des détections précoces lors des présentations audiovisuelle et auditive indiquaient que les participants se basaient sur un indice acoustique présent dans le signal (e.g., l'intonation) afin de détecter plus efficacement cette consonne. Cette identification précoce ne pourrait être que par le biais d'informations extérieures aux stimuli car c'est seulement à partir du gate 90 que le mouvement de fermeture labiale débute. En effet, la séquence débute dans le /a/ qui précède la consonne, avant même que le début de fermeture labiale ne commence.

5.2.5 DESCRIPTION DES STATISTIQUES UTILISEES

5.2.5.1 DETECTIONS CORRECTES

Compte tenu de notre design expérimental, la procédure classique d'analyse de données en *gating*, (à savoir, l'identification des points d'Isolation et de Reconnaissance par individu) ne pouvait être utilisée dans notre étude. Nous ne pouvons en effet pas déterminer ces points par individu, puisque chaque gate n'a été vu qu'une seule fois par chaque participant. Compte tenu de la nature des données, nous avons mené une analyse par t-tests, celle-ci étant le seul outil statistique à notre disposition. Les analyses présentées ici seront donc sommaires, dans le sens où celles-ci ne présentent pas les traitements statistiques finaux.

Concernant les analyses présentées dans ce manuscrit, nous avons déterminé, à partir du pourcentage de détections correctes (DC) du groupe, trois moments d'intérêt (ou « seuil ») pour les présentations auditive, audiovisuelle et visuelle respectivement. Ces seuils correspondaient aux *gates* pour lesquels le pourcentage de DC était d'au moins 10, 50 et 90%. La valeur de *gate* à 10% de DC correspond à une partie du signal où nous sommes sûrs que les participants n'ont pas assez d'information pour répondre correctement. La valeur de *gate* à 50% de DC correspond à une partie du signal où les participants ont assez d'information pour répondre correctement mais le degré d'incertitude concernant l'identité du phonème reste important. La valeur de *gate* à 90% de DC correspond à une partie du signal où nous sommes sûrs que la quantité d'information fournie est assez importante pour la détection. Une fois ces *points* identifiés, et ce, pour chacune des trois modalités de présentation, les analyses ont été menées en trois temps.

(1) Nous avons sélectionné les trois *gates* pour lesquels les pourcentages de DC étaient de 10, 50 et 90% lors de la présentation *auditive.* Ces seuils « auditifs » seront nos références pour les comparaisons statistiques. Ainsi, nous comparons les scores obtenus pour ces trois *gates* dans les deux autres modalités (i.e., audiovisuelle et visuelle). Par exemple, si 50 % de DC sont atteintes au *gate* 100 en modalité auditive, les comparaisons seront réalisées entre les scores obtenues pour ce *gate* (i.e., 100) pour les modalités de présentation audiovisuelle et visuelle seule.

(2) La même opération a été réalisée en prenant comme référence les trois *gates* à partir desquels les scores étaient de 10, 50 et 90% lors de la présentation *audiovisuelle*. Nous avons comparé les scores obtenus pour chacun de ces trois *gates* aux scores obtenus pour les mêmes *gates* dans les deux autres modalités de présentation (i.e., auditive et visuelle).

(3) Et enfin l'analyse a été répétée sur les trois *gates* pour lesquels les pourcentages de détection correcte étaient de 10, 50 et 90% lors de la présentation *visuelle.*
Ces comparaisons seront réalisées séparément pour chaque consonne. Nous présenterons dans un premier temps les résultats obtenus pour la consonne /p/, puis /t/ et enfin /k/. Les comparaisons ont été réalisées avec des T de Student. Le seuil de significativité α a été corrigé par l'application de la correction de Bonferroni. Celui-ci sera significatif à $p< 0.002$.

Ces analyses statistiques seront suivies d'un comparatif inter-consonnes et inter-modalités, ainsi que l'analyse de l'avantage audiovisuel (sur les informations auditive et visuelle seules). Cela nous permettra de comparer les consonnes entre elles à la lumière de leurs caractéristiques articulatoires et/ou auditives. Nous effectuerons également le calcul du gain temporel obtenu pour le seuil de détection de 90% par rapport au *gate* pour lequel l'information auditive sur l'identité de la consone est disponible (i.e., *gate* 370) afin de quantifier l'avantage temporel fourni par l'information visuelle (i.e., les mouvements articulatoires préparatoires) ainsi que l'anticipation éventuelle par les indices acoustiques liés à la coarticulation.

Enfin, nous avons calculé la contribution des modalités auditive et visuelle dans le processus de détection pour chaque *gate*. Cela nous permet de savoir quelle modalité influence le plus la détection en fonction de la durée du signal délivré. Pour ce faire, nous avons comparé l'apport de chaque modalité pour chaque *gate* en soustrayant les taux de DC obtenus lors des présentations auditive et visuelle à ceux de la modalité audiovisuelle.

5.3 RESULTATS

5.3.1.1 TEMPS DE REPONSES

Les patterns obtenus diffèrent en effet beaucoup de nos attentes. Nous avions fait l'hypothèse générale que les temps de réponse allaient diminuer en fonction de la quantité d'information présentée. Autrement dit, nous nous attendions à une diminution progressive de temps jusqu'à atteindre un pallier où le participant serait en mesure de répondre correctement et le plus rapidement possible. Ceci n'est pas ce que nous avons observé. Nous avons globalement constaté des temps de réponse extrêmement variables en début de présentation, lorsque les

participants disposaient de très peu d'informations pour se prononcer quant à l'identité du phonème. Puis, à partir d'un certain point, les temps de réponse se stabilisent autour d'une valeur donnée. Ce pattern de résultats nécessite un traitement de données spécifique et nous n'avons pas pu trouver pour l'instant un modèle statistique satisfaisant pour le faire. C'est pour cette raison que la présentation des temps de réponse sera brève et ne sera pas prise en compte pour l'interprétation des résultats.

5.3.2 POURCENTAGE DE DETECTIONS CORRECTES

5.3.2.1 IMPACT DE LA MODALITE SUR LES PERFORMANCES

5.3.2.1.1 PHONEME /p/

La figure 45 représente les pourcentages de DC moyens obtenus pour chacun des *gates* en fonction de la modalité de présentation (auditive, audiovisuelle, visuelle) pour la séquence /apa/.

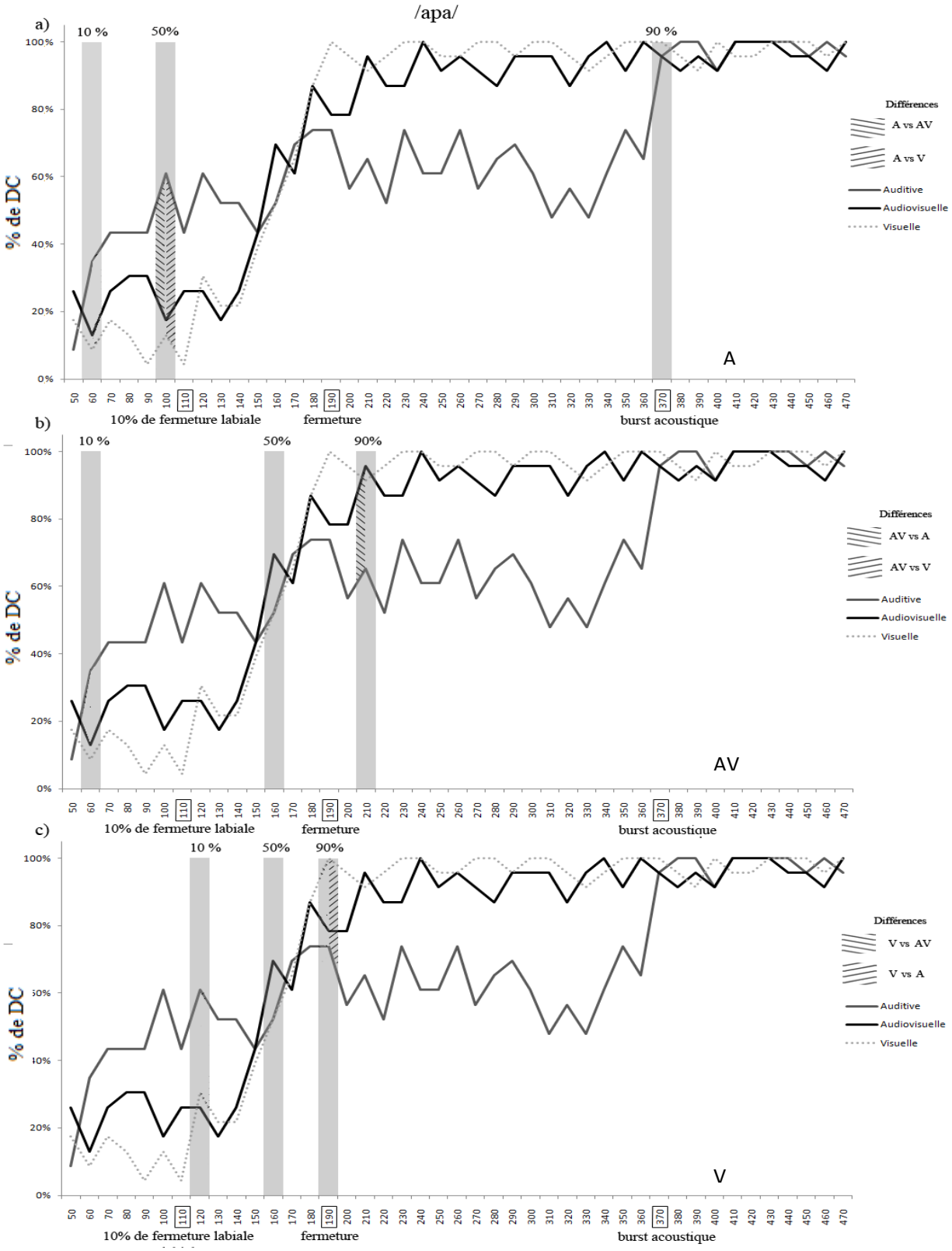

Figure 43. Pourcentage de détections correctes (DC) lors de la présentation de /apa/ pour chaque *gate* en fonction de la modalité de présentation (auditive, audiovisuelle et visuelle). Les indices articulatoires et acoustiques sont indiqués par des encarts sur l'axe des abscisses. Ils indiquent respectivement le début de la fermeture labiale, la fermeture complète et le *burst* acoustique. Les barres grisées représentent les *gates* pour lesquels les pourcentages de DC moyen atteignaient 10, 50 et 90 % lors de la présentation a) auditive, b) audiovisuelle et c) visuelle. Les zones hachurées représentent les différences significatives obtenues par Test de Student en fonction des comparaisons réalisées.

(1) Les « seuils » auditifs de 10, 50 et 90% de DC ont été atteints à partir de 60, 100 et 370 ms respectivement. Nous avons comparé le pourcentage de DC obtenu lors de la présentation *auditive* pour ces *gates* avec le pourcentage de DC obtenu au même *gate* dans les deux autres modalités. Les résultats des tests de Student sont présentés dans le Tableau 1. Les zones de significativité en fonction des comparaisons réalisées sont représentées sur la Figure 45a par des hachures.

/apa/											
Auditive											
10% DC (valeur gate = 60)				50% DC (valeur gate = 100)				90% DC (valeur gate = 370)			
A vs AV		**A vs V**		**A vs AV**		**A vs V**		**A vs AV**		**A vs V**	
t value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value
2.13	0.04	2.6	0.01	**3.8**	**<0,001**	**4.39**	**<0,001**	0.68	0.5	1.44	0.16

Tableau 1. Résultats des Tests de Student effectués sur les pourcentages de DC obtenus en modalité auditive et audiovisuelle (A vs AV) ainsi que auditive et visuelle (A vs V). Les tests ont été réalisés pour les trois *gates* correspondant à 10, 50 et 90% de DC en modalité Auditive. Les valeurs des trois gates sont spécifiées ainsi que les valeurs *t* et *p* associées. Les *p values* notées en gras sont significatives après l'application de la correction du seuil Bonferroni.

Au *gates* 60 et 370 (pour lesquels 10 et 90% de DC sont atteint lors de la présentation auditive), aucune différence n'est induite par la Modalité de présentation. Pour ces deux *gates*, les scores de détection sont donc équivalents quelque soit la Modalité de présentation. Au *gate* 100, pour lequel les performances de détection auditive atteignent 50%, le pourcentage de DC est plus important lors de la présentation auditive ($M_{auditive}$ = 60% ; SD = 10) qu'audiovisuelle ($M_{audiovisuelle}$ =18% ; SD = 8 ; ($t_{(22)}$ = 3.8, p < .001) ou visuelle ($M_{Visuelle}$ = 13% ; SD = 7 ; ($t_{(22)}$ = 4.39, p < .001).

(2) En suivant le même schéma nous avons réalisé les mêmes comparaisons en nous basant sur les trois *gates* pour lesquelles le pourcentage de DC était de 10, 50 et 90 % lors de la présentation *audiovisuelle* (i.e., les *gates* 60, 160 et 210 respectivement) et les avons comparé au scores obtenus pour ces mêmes *gates* lors des présentations auditive et visuelle. Les résultats des tests de Student sont présentés dans le Tableau 2. Les zones de significativité en fonction de la comparaison réalisée sont représentées sur la Figure 45b par des hachures.

/apa/											
Audiovisuelle											
10% DC (valeur gate = 60)				50% DC (valeur gate = 160)				90% DC (valeur gate = 210)			
AV vs A		**AV vs V**		**AV vs A**		**AV vs V**		**AV vs A**		**AV vs V**	
t value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value
2.13	0.04	1.01	0.32	1.62	0.11	1.62	0.11	**3.18**	**<0,001**	1.1	0.28

Tableau 2. Résultats des Tests de Student effectués sur les pourcentages de DC obtenus en modalité audiovisuelle et auditive (AV vs A) ainsi que audiovisuelle et visuelle (AV vs V). Les tests ont été réalisés pour les trois *gates* correspondant à 10, 50 et 90% de DC en modalité audiovisuelle. Les valeurs des trois gates sont spécifiées ainsi que les valeurs *t* et *p* associées. Les *p values* notées en gras sont significatives après l'application de la correction du seuil Bonferroni.

Lorsque le pourcentage de DC atteint 10 % et 50% lors de la présentation audiovisuelle, aucune différence significative n'est observée entre les modalités. Les comparaisons nous permettent de constater que la courbe audiovisuelle atteint 90 % de détection bien avant la courbe auditive, puisque lorsque les scores en audiovisuelle atteignent 95% (SD = 4) (*gate* 210), les scores en auditif sont largement inférieurs avec une moyenne de détection de 65 % (SD = 10, $t_{(22)} = 3.18$, $p < .001$). Aucune différence n'est observée entre les scores audiovisuel et visuel ($t_{(22)} = 1.1$, $p = .28$).

(3) Enfin, les comparaisons ont été réalisées sur la base des pourcentages de DC obtenus lors de la présentation *visuelle*. Les résultats des tests de Student sont présentés dans le Tableau 3. Les zones de significativité en fonction de la comparaison réalisée sont représentées sur la Figure 45c par des hachures.

/apa/											
Visuelle											
10% DC (valeur gate = 120)				50% DC (valeur gate = 160)				90% DC (valeur gate = 190)			
V vs AV		**V vs A**		**V vs AV**		**V vs A**		**V vs AV**		**V vs A**	
t value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value
0.9	0.37	2.52	0.01	1.62	0.11	0.68	0.5	2.88	0.008	**3.22**	**0.003**

Tableau 3. Résultats des Tests de Student effectués sur les pourcentages de DC obtenus en modalité visuelle et audiovisuelle (V vs AV) ainsi que visuelle et auditive (V vs A). Les tests ont été réalisés pour les trois *gates* correspondant à 10, 50 et 90% de DC en modalité visuelle. Les valeurs des trois *gates* sont spécifiées ainsi que les valeurs *t* et *p* associées. Les *p values* notées en gras sont significatives après l'application de la correction du seuil Bonferroni.

Les comparaisons indiquent que lorsque la modalité visuelle permet d'atteindre 90 % de DC, les scores auditifs sont tendantiellement (suite à la correction de seuil de Bonferroni) inférieurs avec une moyenne de 7% (SD = 9 ; $t_{(22)} = 3.22$, $p = .003$).

5.3.2.1.2 PHONEME /t/

La figure 46 représente les pourcentages de DC moyens obtenus pour chacun des *gates* en fonction de la modalité de présentation (auditive, audiovisuelle, visuelle) pour la séquence /ata/.

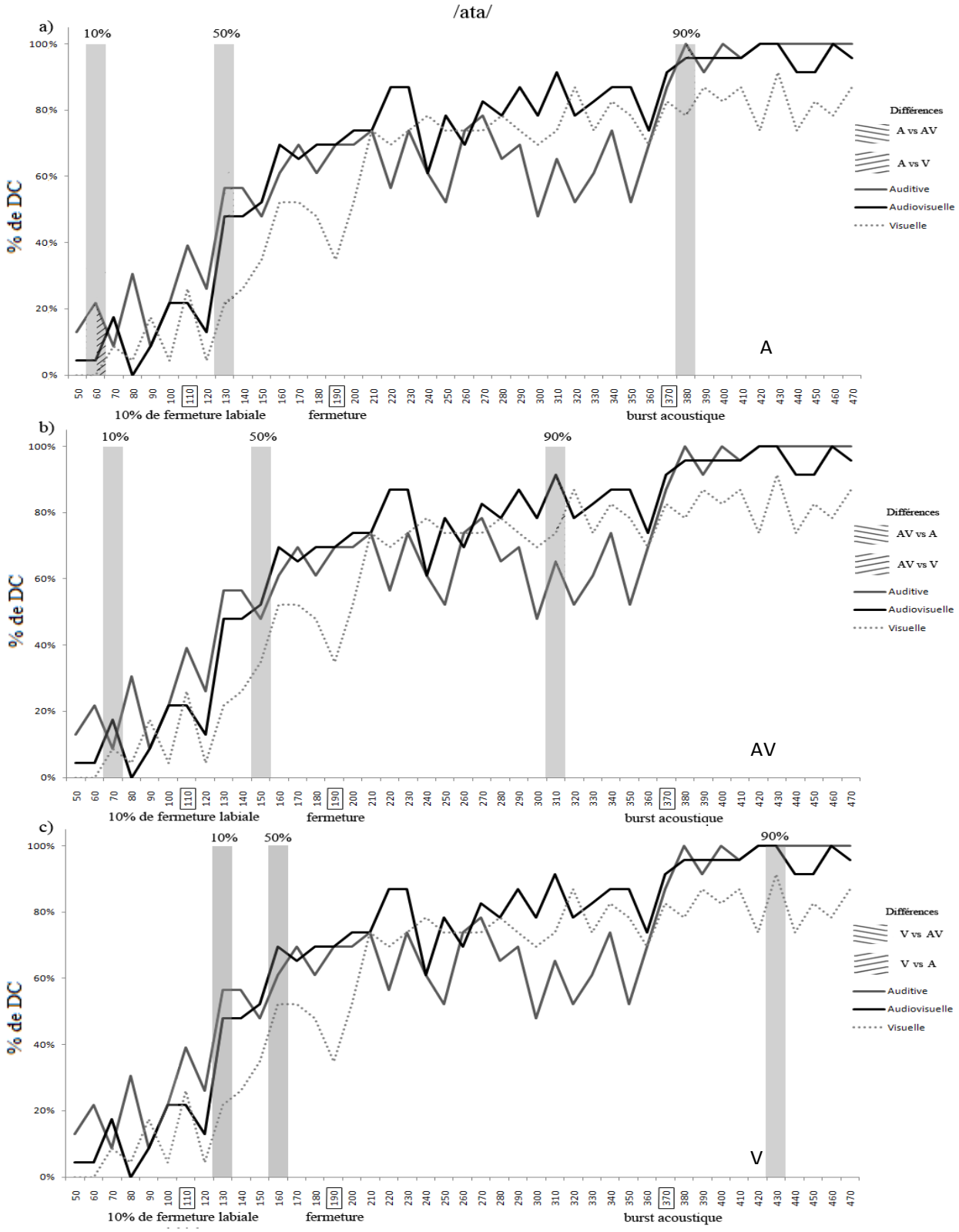

Figure 44. Pourcentage de détections correctes (DC) lors de la présentation de /ata/ pour chaque *gates* en fonction de la modalité de présentation (Auditive, Audiovisuelle et Visuelle). Les moments clé du signal sont indiqués par des encarts sur l'axe des abscisses qui indiquent respectivement le début de la fermeture labiale, la fermeture complète (qui équivaut pour le /t/ au point de stabilisation de la fermeture) et le *burst* acoustique. Les barres grisées représentent les valeurs de *gates* pour lesquelles les pourcentages de DC moyen atteignaient 10, 50 et 90 % lors de la présentation a) auditive, b) audiovisuelle et c) visuelle. Les zones hachurées représentent les différences significatives obtenues par Test de Student en fonction des comparaisons réalisées.

(1) Nous avons comparé le pourcentage de DC obtenu lors de la présentation auditive pour les *gates* 60, 130 et 380 correspondant respectivement à 10, 50 et 90% de DC avec le pourcentage de DC obtenu au même moment dans les deux autres modalités. Les résultats des tests de Student sont présentés dans le Tableau 4. La zone de significativité en fonction de la comparaison réalisée est représentée sur la Figure 46a par des hachures.

/ata/											
Auditive											
10% DC (valeur gate = 60)				50% DC (valeur gate = 130)				90% DC (valeur gate = 380)			
A vs AV		**A vs V**		**A vs AV**		**A vs V**		**A vs AV**		**A vs V**	
t value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value
2.16	0.04	**2.88**	**<0,001**	1.1	0.28	1.44	0.007	1.44	1.16	2.88	0.008

Tableau 4. Résultats des Tests de Student effectués sur les pourcentages de DC obtenus en modalité auditive et audiovisuelle (A vs AV) ainsi que auditive et visuelle (A vs V). Les tests ont été réalisés pour les trois *gates* correspondant à 10, 50 et 90% de DC en modalité Auditive. Les valeurs des trois gates sont spécifiées ainsi que les valeurs *t* et *p* associées. Les *p values* notées en gras sont significatives après l'application de la correction de seuil de Bonferroni.

Pour la consonne /t/, nous pouvons constater que lorsque les pourcentages de DC dépassent le seuil de 10% en modalité auditive, le score dans la condition visuelle est inférieur avec une moyenne de 0% (SD = 0 ; $t_{(22)} = 2.88$, $p < .001$) contre de 21% (SD = 8) lors de la présentation auditive. Aucune différence n'atteignent la significativité après l'application de la correction de Bonferroni.

(2) En suivant le même principe, nous avons réalisé les mêmes comparaisons en nous basant sur les *gates* pour lesquelles les pourcentages de DC étaient de 10, 50 et 90 % lors de la présentation audiovisuelle. Ils correspondent aux *gates* 70, 150 et 310. Les résultats des tests de Student sont présentés dans le Tableau 5.

/ata/											
Audiovisuelle											
10% DC (valeur gate = 70)				50% DC (valeur gate = 150)				90% DC (valeur gate = 310)			
AV vs A		**AV vs V**		**AV vs A**		**AV vs V**		**AV vs A**		**AV vs V**	
t value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value
1.33	0.16	1.33	0.19	0.88	0.38	1.6	0.12	2.6	0.01	1.96	0.06

Tableau 5. Résultats des Tests de Student effectués sur les pourcentages de DC obtenus en modalité audiovisuelle et auditive (AV vs A) ainsi que audiovisuelle et visuelle (AV vs V). Les tests ont été réalisés pour les trois *gates* correspondant à 10, 50 et 90% de DC en modalité audiovisuelle. Les valeurs des trois gates sont spécifiées ainsi que les valeurs *t* et *p* associées. Les *p values* notées en gras sont significatives après l'application de la correction de seuil de Bonferroni.

Les scores obtenus pour les *gates* 70, 150 et 310 (i.e., les *gate* pour lesquels 10, 50 et 90% de DC était obtenue en modalité audiovisuelle) sont équivalents entre les modalités.

(3) Enfin, les comparaisons ont été effectuées de nouveau en utilisant comme *gates* de référence ceux pour lesquels a présentation visuelle permettait 10, 50 et 90% de DC (i.e., les *gates* 100, 130 et 430 respectivement. Les résultats des comparaisons sont présentés Tableau 6. Les différence significatives observées sont reportées sur la Figure 46c.

/ata/											
Visuelle											
50% DC (valeur gate = 130)				50% DC (valeur gate = 160)				90% DC (valeur gate = 430)			
V vs AV		**V vs A**		**V vs AV**		**V vs A**		**V vs AV**		**V vs A**	
t value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value
1.1	0.28	0.68	0.5	1.62	0.11	1.1	0.28	1.85	0.07	1.85	0.07

Tableau 6. Résultats des Tests de Student effectués sur les pourcentages de DC obtenus en modalité visuelle et audiovisuelle (V vs AV) ainsi que visuelle et auditive (V vs A). Les tests ont été réalisés pour les trois *gates* correspondant à 10, 50 et 90% de DC en modalité Auditive. Les valeurs des trois *gates* sont spécifiées ainsi que les valeurs *t* et *p* associées. Les *p values* notées en gras sont significatives après l'application de la correction de seuil de Bonferroni.

Aucune différence significative n'est observée.

5.3.2.1.3 PHONEME /k/

La Figure 47 représente les pourcentages de DC moyens obtenus pour chacun des *gates* en fonction de la modalité de présentation (auditive, audiovisuelle, visuelle) pour la séquence /aka/.

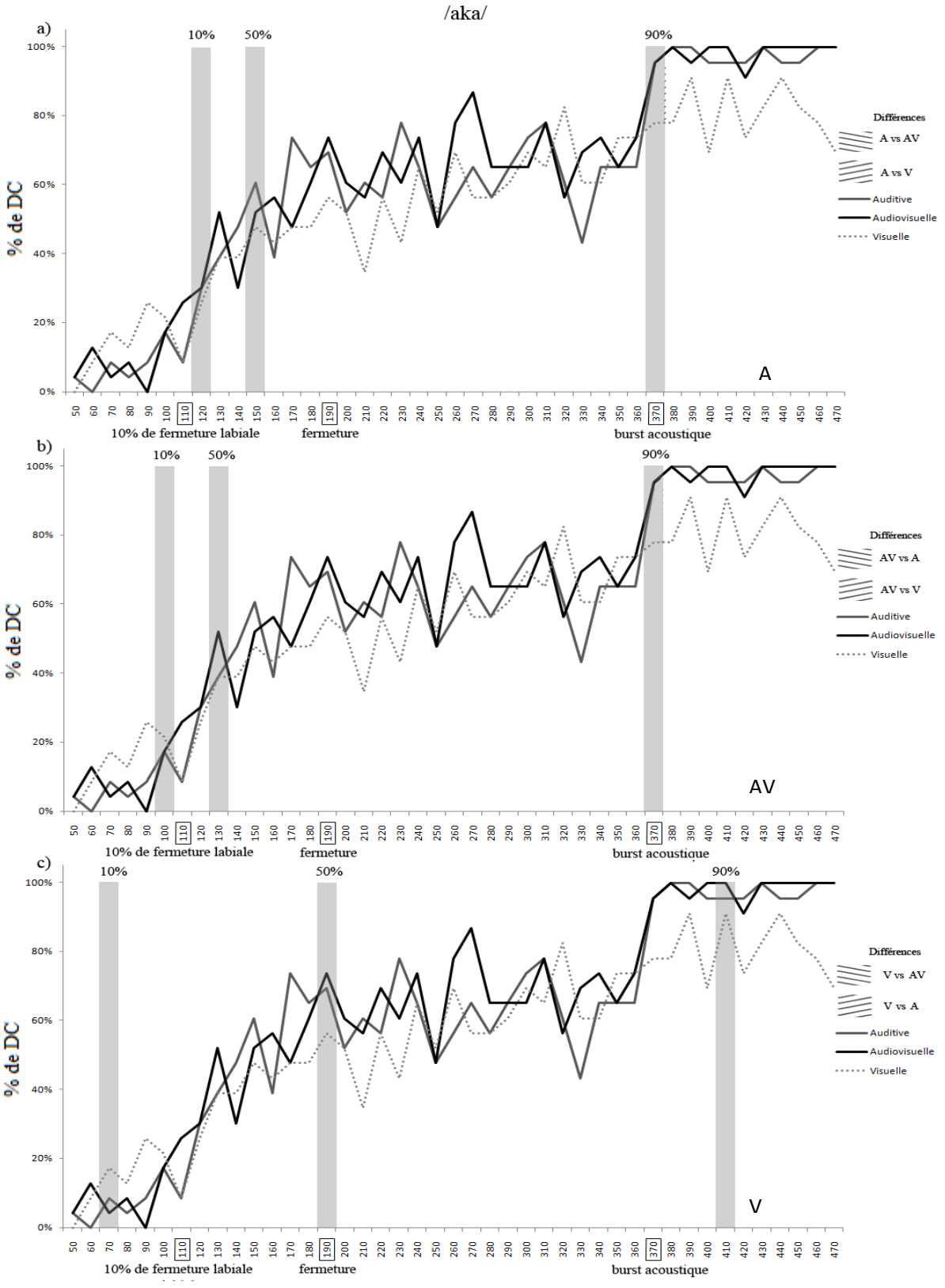

Figure 45. Pourcentage de détections correctes (DC) lors de la présentation de /aka/ pour chaque *gates* en fonction de la modalité de présentation (Auditive, Audiovisuelle et Visuelle). Les moments clé du signal sont indiqués par des encarts sur l'axe des abscisses qui indiquent respectivement le début de la fermeture labiale, la fermeture complète (qui équivaut pour le /t/ au point de stabilisation de la fermeture) et le *burst* acoustique. Les barres grisées représentent les valeurs de *gates* pour lesquelles les pourcentages de DC moyen atteignaient 10, 50 et 90 % lors de la présentation a) Auditive, b) audiovisuelle et c) visuelle. Les zones hachurées représentent les différences significatives obtenues par Test de Student en fonction des comparaisons réalisées.

(1) Lors des comparaisons effectuées entre les modalités pour la séquence /aka/, aucune des comparaisons n'a révélé de différences significatives, quelque soit le *gate* considéré. Les résultats des tests de Student sont présentés dans les Tableaux 7, 8 et 9.

/aka/											
Auditive											
10% DC (valeur gate = 120)				50% DC (valeur gate = 150)				90% DC (valeur gate = 370)			
A vs AV		**A vs V**		**A vs AV**		**A vs V**		**A vs AV**		**A vs V**	
t value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value
0.68	0.5	0.9	0.37	1.1	0.28	1.34	0.9	0.68	0.5	2.16	0.04

Tableau 7. Résultats des Tests de Student effectués sur les pourcentages de DC obtenus en auditive et audiovisuelle (A vs AV) ainsi que auditive et visuelle (A vs V). Les tests ont été réalisés pour les trois *gates* correspondant à 10, 50 et 90% de DC en modalité Auditive. Les valeurs des trois gates sont spécifiées ainsi que les valeurs *t* et *p* associées. Les *p values* notées en gras sont significatives après l'application de la correction de seuil de Bonferroni.

/aka/											
Audiovisuelle											
10% DC (valeur gate = 100)				50% DC (valeur gate = 130)				90% DC (valeur gate = 370)			
AV vs A		**AV vs V**		**AV vs A**		**AV vs V**		**AV vs A**		**AV vs V**	
t value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value
0.68	0.5	0.93	0.35	1.34	0.19	1.34	0.19	0.68	0.5	2.16	0.04

Tableau 8. Résultats des Tests de Student effectués sur les pourcentages de DC obtenus en modalité audiovisuelle et auditive (AV vs A) ainsi que audiovisuelle et visuelle (AV vs V). Les tests ont été réalisés pour les trois gates correspondant à 10, 50 et 90% de DC en modalité Audiovisuelle. Les valeurs des trois gates sont spécifiées ainsi que les valeurs *t* et *p* associées. Les *p values* notées en gras sont significatives après l'application de la correction de seuil de Bonferroni.

/aka/											
Visuelle											
10% DC (valeur gate = 70)				50% DC (valeur gate = 190)				90% DC (valeur gate = 410)			
V vs AV		**V vs A**		**V vs AV**		**V vs A**		**V vs AV**		**V vs A**	
t value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value	*t* value	*p* value
1.82	0.08	1.33	0.19	1.65	0.11	1.36	0.18	1.85	0.07	1.1	0.28

Tableau 9. Résultats des Tests de Student effectués sur les pourcentages de DC obtenus en modalité visuelle et audiovisuelle (V vs AV) ainsi que visuelle et auditive (V vs A). Les tests ont été réalisés pour les trois *gates* correspondant à 10, 50 et 90% de DC en modalité Visuelle. Les valeurs des trois *gates* sont spécifiées ainsi que les valeurs *t* et *p* associées. Les *p values* notées en gras sont significatives après l'application de la correction de seuil de Bonferroni.

5.3.2.2 IMPACT DE LA MODALITE SUR LES SEUILS DE DETECTION EN FONCTION DE LA CONSONNE

Le tableau 10 présente un récapitulatif des *gates* d'intérêt (pour lequels 10, 50 et 90% de DC sont atteints) en fonction de la Consonne à détecter et de la Modalité de présentation. Il présente également l'avantage audiovisuel (A-AV et V-AV) ainsi que le bénéfice temporel de détection (i.e., différence entre le *gate* où 90% de DC est atteint et le *gate* où le *burst* est relâché).

	Récapitulatif			Avantages				
	10% A	10% AV	10% V	A - AV	V - AV			
/p/	60	60	120	0	60			
/t/	60	70	100	-10	30			
/k/	120	100	70	20	-30			
	50% A	50% AV	50% V	A - AV	V - AV			
/p/	100	160	160	-60	0			
/t/	130	150	130	-20	-20			
/k/	150	130	190	20	60	Bénéfice/acoustique		
	90% A	90% AV	90% V	A - AV	V - AV	A	AV	V
/p/	370	210	190	160	-20	0	160	180
/t/	380	310	430	70	120	-10	60	-60
/k/	370	370	410	0	40	0	0	-40

Tableau 10. Récapitulatif des différents *gates* d'intérêt (i.e. 10, 50 et 90% de DC) en fonction de la consonne à détecter pour les modalités de présentation audiovisuelle (AV), auditive (A) et visuelle (V). La colonne "Avantage" correspond à l'avantage audiovisuelle observé par rapport aux présentations auditive (*gate* A - *gate* AV) et visuelle (*gate* V - *gate* AV) pour chacun des pourcentages de DC d'intérêt. Les valeurs positives indiquent donc un avantage temporel de détection en modalité audiovisuelle. La colonne "Bénéfice/acoustique" correspond au bénéfice temporel observé entre le *gate* auquel l'information acoustique sur l'identité de la consonne est disponible et le *gate* pour lequel 90% de DC est atteint (370-*gate* 90%).

D'un point de vue général nous pouvons constater que les profils d'évolution temporelle entre /k/ et /t/ sont plus similaires entre eux qu'ils ne le sont de /p/. Ce dernier atteint en effet 90% de DC bien avant les deux autres consonnes puisque cette détection s'opère aux *gates* 210 et 190 pour les modalités audiovisuelle et visuelle respectivement alors ces scores ne sont atteints par /t/ et /k/ au mieux qu'au *gate* 310 lors de la présentation audiovisuelle et 410 lors de la présentation visuelle seule. La consonne /p/ est donc détectée à hauteur de 90% au moins 100 ms avant /t/ et /k/ lors d'une présentation audiovisuelle et 220 ms lors de la présentation visuelle. Nous constatons d'ailleurs que le *gate* auquel la consonne /p/ est détectée visuellement à hauteur de 90% correspond à la fermeture labiale.

L'analyse de l'avantage fourni par la présentation audiovisuelle par rapport aux présentations auditive et visuelle nous renseigne sur l'informativité de chaque canal. Pour la détection du phonème /p/, nous pouvons constater qu'alors que 50% de DC est atteint au *gate* 100 lors de la présentation auditive, ce pourcentage n'est atteint que 60 ms plus tard en présentation audiovisuelle. L'écart se creuse par la suite avec un pourcentage de DC de 90 % atteint au *gate* 210 soit, 270 ms plus tôt que lors d'une présentation auditive. Les informations visuelles semblent également fournir un avantage puisque 90% de DC sont atteint 20 ms plus tôt lors de la présentation visuelle par rapport à la présentation audiovisuelle. La consonne /t/, dont l'articulation est moins visible que /p/ bénéficie tout de même d'un avantage audiovisuel

de 70 ms pour le *gate* permettant 90% de DC. Cependant, la présentation des informations visuelles seules ne permet d'atteindre 90% de DC 120 ms plus tard que dans le cadre de la présentation bimodale. Enfin, un très faible avantage est obtenu pour la consonne vélaire /k/ avec un gain maximun de 20 ms permis par la présentation audiovisuelle par rapport à la présentation auditive seule.

L'observation du bénéfice obtenu en terme de détection, c'est-à-dire la différence entre le *gate* pour lequel l'information acoustique sur l'identité de la voyelle est disponible (i.e., 360) et le *gate* pour lequel le pourcentage de DC atteint 90% nous indique que toutes les consonnes sont identifiées lors de la présentation auditive au moment où l'information auditive sur l'identité de la consonne est disponible ou 10 ms plus tard dans le cas de /t/. La présentation audiovisuelle permet quant à elle une gradation du bénéfice en fonction de la saillance visuelle des consonnes. La consonne /p/ est celle qui bénéficie de l'avantage temporel le plus important puisque 90% de DC est atteint 160 ms avant le *burst*, suivi de /t/ qui est détecté 60 ms avant le *burst*. Enfin la détection de /k/ n'est pas avantagé par la présentation multimodale puisque la consonne est détectée au même moment que lorsque seule l'information auditive est disponible.

5.3.3 CONTRIBUTION DIFFERENTIELLE DE CHAQUE MODALITE AU COURS DU TEMPS

La contribution de chacune des modalités a été calculée pour les trois consonnes (Figure 48). En soustrayant les résultats obtenus lors des présentations unimodales aux scores obtenus lors de la présentation bimodale, nous pouvons observer d'une part quelle modalité est la plus informative en fonction du temps et d'autre part l'avantage audiovisuel.

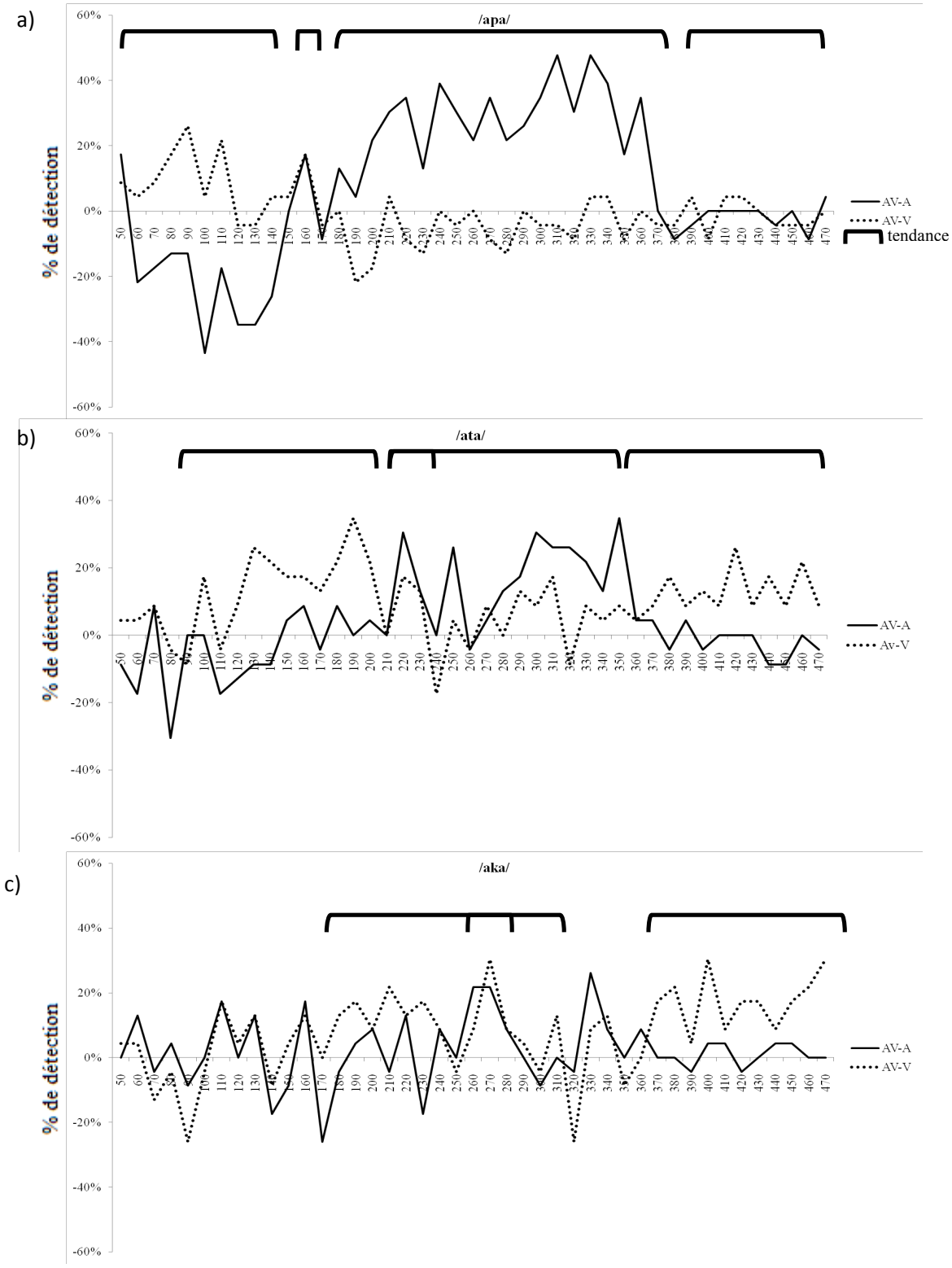

Figure 46. Contribution des modalités auditive et visuelle dans le processus de détection des consonnes /p/ (a), /t/ (b) et /k/ (c) obtenu par soustraction des scores unimodaux aux scores bimodaux (AV-A et AV-V) pour chaque *gate.* Les délimitations placées au-dessus des courbes représentent des patterns de contribution similaires (intra-délimitation) ou dissimilaires (inter- délimitations).

5.3.3.1.1.1 PHONEME /p/

Quatres tendances sont observées lors de la présentation que la consonne /p/ (Figure 48a) :

Dans un premier temps, nous observons une double tendance avec des scores AV-A négatifs et AV-V positifs. L'avantage des scores obtenus lors de la présentation audiovisuelle sur la présentation visuelle est principalement dû aux informations auditives présentes dans le signal. Cela est coroboré par le fait que la résultante de AV-A est dans le négatif, indiquant une infériorité des scores obtenus lors de la présentation audiovisuelle par rapport à la présentation auditive. Pour cette partie du signal, les informations auditives seules sont donc les plus informatives et ne sont pas utilisées de la même façon lors de la présentation auditive et audiovisuelle. Les participants se basent principalement sur l'information auditive, les informations visuelles présentes lors de la présentation audiovisuelle semblent gêner la détection.

Autour de 160 ms de présentation, nous observons un *pattern intégratif* : les résultantes de AV-A et AV-V sont positives, ce qui indique une supériorité des scores audiovisuels à la fois sur la présentation auditive et sur la présentation visuelle. Deux t-tests ont été effectués afin de savoir si la différence entre les scores AV et A et AV et V était significative. Si les scores audiovisuels sont supérieurs aux deux autres présentations, cela indique l'accès à des indices complémentaires fournis lors de la présentation bimodale. Ces tests ont révélé qu'alors que la présentation audiovisuelle permet 17% de DC supplémentaires par rapport aux présentations visuelle et auditive, mais cette différence n'est pas significative ($t(22) = 1.62$, $p = .11$).

Par la suite, de 180 ms à 370 ms de présentation (i.e., les *gates* allant de la fermeture labiale au relâchement du *burst*), nous observons que la résultante de AV-A est positive alors que AV-V donne des résultats nuls. Les résultats obtenus lors de la présentation audiovisuelle sont donc majoritairement dûs aux informations visuelles puisque les scores AV et V sont similaires. Il y a donc un avantage audiovisuel, celui-ci étant dû aux informations visuelles puisque lorsque l'on soustrait les scores auditifs aux scores audiovisuels, les scores restent positifs. Ce ne sont donc pas les informations auditives qui sont principalement utilisées pour la détection de /p/.

Enfin, après le *burst*, aucune modalité ne prend le dessus pour la détection, toutes les informations étant informatives après 370 ms quelle que soit la modalité de présentation.

5.3.3.1.1.2 PHONEME /t/

Quatres tendances sont observées lors de la présentation que la consonne /t/ (Figure 48b) :

Dans un premier temps (i.e., de 100 à 210 ms), la tendance est à une supériorité des informations auditives puisque la résultante de AV-V est positive. Le signal visuel n'explique donc pas a lui seul les résultats obtenus lors de la présentaiton audiovisuelle. Les informations auditives sont les plus informatives lorsque cette quantité de signal est disponible.

Par la suite, un pattern intégratif est observé à 220 ms puisque les résultantes de AV-V et AV- A sont positives. Les comparaisons statistiques effectuées entre les modalités pour ce *gate* nous indiquent que le score audiovisuel est alors significativement supérieur au score auditif ($t(22) = 2.78$, $p < .016$) à hauteur de 30% et tendantiellement différent du score obtenu lors de la présentation visuelle ($t(22) = 1.83$, $p = 0.07$) avec un avantage de 17%.

Le pattern AV-V passe par la suite par une phase neutre (les scores en audiovisuelle sont donc fortement liés aux indices visuels) durant laquelle la résultante de AV-A est positive. Nous observons donc ici un renversement de la tendance précédemment observée de 110 à 210 ms, en faveur des informations visuelles. En effet, si l'on soustrait les scores auditifs aux scores audiovisuels, la tendance est toujours positive ce qui indique l'utilisation d'indices visuels plutôt qu'auditifs lors de la présentation audiovisuelle.

Dans la suite du signal et contrairement à /p/, la détection se fait sur la base des informations auditives lorsque plus de 370 ms sont disponibles puisque la résultante de AV-V est positive alors que AV-A est nulle.

5.3.3.1.1.3 PHONEME /k/

Deux, voire trois tendances sont observées lors de la présentation que la consonne /k/ (Figure 48c) :

Pour la consonne /k/, on constate que les informations visuelles sont peu informatives car elles ne permettent que dans de rare cas des performances supérieures à la présentation audiovisuelle. De 170 à 240 ms, les informations auditives sont plus informatives que les informations visuelles, nous indiquant que les informations auditives sont utilisées pour la détection.

Des résultats tendantiellement supérieurs aux scores observés lors des présentations unimodales sont obtenus lors de la présentation audiovisuelle reflétant une intégration multimodale à 270 ms avec des résultats 21 % supérieurs en présentation audiovisuelle par rapport à la présentation auditive ($t(22) = 2.13$, $p = .04$) et de 30% supérieurs aux scores visuels ($t(22) = 2.78$, $p < .016$).

Enfin, tout comme pour /t/, lorsque les *gates* supérieurs à 370 sont dévoilés, les réponses se basent préférentiellement sur les informations auditives.

5.4 TEMPS DE REPONSE

La figure 49 représente l'évolution des Temps de réponse au cours du temps en fonction de la Modalité de présentation (auditive, audiovisuelle, visuelle) et de la Consonne à détecter (/p/, /t/, /k/).

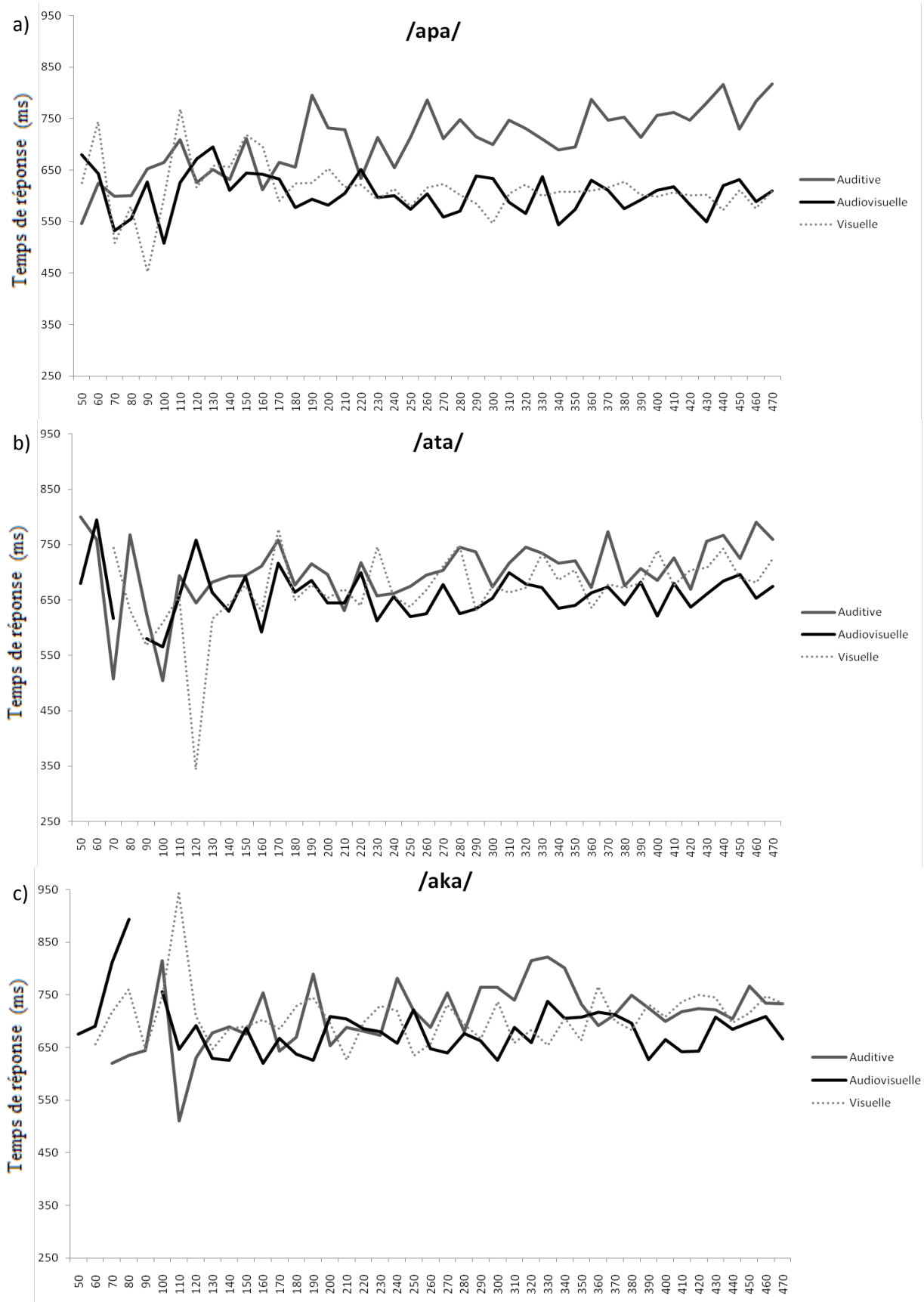

Figure 47. Temps de réponses moyens observés en fonction de la Modalité de présentation (Auditive, Audiovisuelle et Visuelle) et du Gate pour les consonnes /p/ (a), /t/ (b) et /k/ (c).

La moyenne des temps de réponse pour /apa/ est de 705 ms (SD = 13) lors de la présentation auditive, 604 ms (SD = 8) lors de la présentation audiovisuelle et de 613 ms (SD = 10) lors de la présentation visuelle.

La moyenne des temps de réaction pour /ata/ est de 701 ms (SD = 12) lors de la présentation auditive, 659 ms (SD = 8) lors de la présentation audiovisuelle et de 669 ms (SD = 14) lors de la présentation visuelle.

La moyenne des temps de réaction pour /aka/ est de 717 ms (SD = 14) lors de la présentation auditive, 684 ms (SD = 10) lors de la présentation audiovisuelle et de 708 ms (SD = 10) lors de la présentation visuelle.

De manière tout à fait descriptive, nous pouvons constater que les temps de réponses sont équivalents entre les consonnes lors de la présentation auditive. Nous observons également des temps de réponse inférieurs lors de la présentation audiovisuelle par rapport à la présentation auditive, avec une augmentation croissante des temps de réponses avec la diminution de la saillance visuelle.

5.5 DISCUSSION

L'étude présentée ici avait pour but de quantifier de manière fine l'apport de chacune des modalités pour la détection des consonnes plosives bilabiale /p/, dentale /t/ et alvéolaire /k/. Un paradigme de *gating on-line* a été utilisé dans lequel le participant était exposé à des séquences plus ou moins longues des trois types de stimuli (/apa/, /ata/ et /aka/), en modalité (1) audiovisuelle, durant laquelle il pouvait se baser à la fois sur le signal acoustique et sur les mouvements oro-faciaux pour répondre, (2) auditive seule et (3) visuelle seule. Dans chacune de ces conditions, le participant était invité à détecter successivement les consonnes /p/, /t/ et /k/ dans trois listes identiques qui contenaient tous les *gates* des trois séquences. Les participants ne devaient répondre que lorsqu'ils percevaient la consonne d'intérêt et ne rien faire dans les cas où (1) la séquence était trop courte pour que l'information sur l'identité de la consonne ne soit disponible ou (2) s'ils percevaient une autre consonne que la consonne cible.

Quatre indices présents dans le signal sont particulièrement importants et doivent être gardés à l'esprit :

- le début de la fermeture labiale (i.e., 10% de fermeture) apparaît visuellement au *gate* 110 ;
- la fin acoustique de la première voyelle correspond au *gate* 170 ;

(la fermeture labiale totale dans le cas de /p/ correspond au *gate* 190) ;

- une période de silence s'étend de 170 à 360 ms ; et
- le relâchement du *burst* est audible au *gate* 370.

Nous avons fait l'hypothèse qu'un gain audiovisuel serait observé quand les deux modalités sont complémentaires, celui-ci se manifestant par une détection plus précoce ainsi que des performances plus importantes lors de la présentation audiovisuelle. Ce gain devrait disparaître quand l'une ou l'autre des modalités contient assez d'informations pour permettre la détection. Le gain audiovisuel serait donc modulé par la saillance visuelle (Arnal et al., 2009; van Wassenhove et al., 2005) et auditive des phonèmes. Les consonnes articulées à l'avant du conduit vocal devraient bénéficier d'un large bénéfice audiovisuel, mais également d'un fort impact de la modalité visuelle dans les premiers *gates* (Jesse & Massaro, 2010) alors que celles articulées à des places plus en arrière devraient amener une augmentation graduelle du taux d'identification.

5.5.1 SEUIL DIFFERENTIEL DE DETECTION EN MODALITE AUDITIVE : COMPARAISON INTER-CONSONNES

Les seuils de 50% de DC sont obtenus respectivement aux *gates* 100, 160 et 150 respectivement pour /p/, /t/ et /k/ lors de la présentation auditive. Les scores auditifs stagnent autour de 65-70% de DC à partir du *gate* 170 quelque soit la consonne considérée et ce, jusqu'à l'arrivée du *burst*. L'augmentation des performances de 0 à 50% de DC est dûe au fait que lors de la présentation auditive, les participants détectent les transitions formantiques 30 à 40 ms avant la fermeture labiale (Libermann, 1996) ce qui leur permet d'atteindre environ 60% de DC. Cependant, ces indices ne permettent pas d'assurer la détection à 100% car les consonnes que nous avons utilisées sont non voisées. Les transitions formantiques sont dans ce cas moins audibles comparées aux consonnes plosives voisées. Le pourcentage de DC décrit donc un plateau durant la période de *silence* (i.e., *gate* 170 à 370) qui suit les transitions formantiques. En effet, lors de l'articulation d'une consonne plosive, contrairement aux fricatives par exemple, l'occlusion du conduit vocal (qui précède la plosion) entraîne une période de silence. Celle-ci ne permet pas d'accumuler d'information auditive quant à l'identité de la consonne. Les résultats restent donc stables durant cette période. Ce plateau, ne s'observait pas chez Jesse & Massaro (2010) puisque leur nombre de *gates* était très restreint. Dans leur expérience, ils proposaient aux participants d'identifier des mots qui étaient

découpés en 6 *gates*. Par exemple, les participants devaient identifier le mot « *cash* » dans la séquence « *a cash* » (les participants avaient pour consigne d'ignorer le déterminant). Durant le premier *gate*, seul les mouvements préparatoires étaient disponibles. A la fin du troisième *gate*, la première consonne était entièrement dévoilée. Ils observèrent une augmentation progressive des détections basées sur l'information auditive due à une accumulation de signal au cours des *gates*. Dans notre cas, seules les transitions formantiques, disponibles aux *gates* 140-170 permettaient d'accumuler un peu d'information concernant l'identité de la consonne. Il faut donc attendre le *gate* 370, et le relâchement du *burst* pour que de l'information puisse de nouveau être accumulée afin de permettre l'identification de la consonne, et ainsi faire passer les scores de 70 à 90% de DC. Lorsque les informations auditives sur le *burst* acoustique sont disponibles (*gate* 370), les participants sont en mesure de détecter la consonne. Nous n'observons pas d'anticipation auditive puisque les consonnes étudiées, des plosives, dans un contexte vocalique neutre, ne donnent lieu qu'à que peu de coarticulation.

Enfin, contrairement au pattern observé pour /p/ lors du calcul des contributions de chaque modalité aux scores multimodaux, la détection de /k/ et /t/ continue à s'appuyer sur les indices auditifs lorsque plus de signal est disponible. On constate qu'après le *gate* 370, où le *burst* acoustique est relâché, les détections se font majoritairement sur la base des informations auditives (courbe AV-V > 0), celles-ci étant plus informatives pour discriminer ces deux consonnes.

Nous observons une contribution ponctuelle des informations auditives dans le processus de reconnaissance de plosives non voisées qui ne varient qu'en terme de place d'articulation (caractéristique peu audible mais parfois utilisée pour identifier les consonnes dont la place d'articulation est plus à l'arrière du conduit vocal ; Jesse & Massaro, 2010). Le premier pic est atteint lorsque les transitions formantiques sont dévoilées. Il faudra attendre le *burst* pour que le cumul d'informations sur l'identité de la consonne soit suffisant et permettre l'identification.

5.5.2 BENEFICE AUDIOVISUEL

Dans notre étude, et contrairement aux résultats précédemment obtenus dans la littérature (Cathiard, 1994 ; Jesse & Massaro, 2010 ; Munhall & Tohkura, 1998 ; Smeele, 1994), aucun avantage audiovisuel global n'est obtenu. L'association de l'information auditive et visuelle n'amène pas nécessairement plus de réponses correctes que la présentation des indices

unimodaux. En effet, si l'une ou l'autre des deux sources d'informations fournit des indices importants sur l'identité de la consonne (e.g., fermeture labiale du /p/), cette information sera utilisée lors de la présentation unimodale et bimodale de la même façon. Comme Grant et collaborateurs (Grant et al., 1998 ; Grant & Walden, 1996), lorsque les modalités auditive et visuelle fournissent les mêmes indices liés à l'identité du phonème, un faible avantage est observé en présentation audiovisuelle. *A contrario*, l'avantage devrait être important quand les modalités fournissent des informations différentes/complémentaires. Une anticipation sera permise si les indices visuels précèdent le flux acoustique (Chandrasekaran et al., 2009) et s'ils sont assez saillants pour pouvoir être utilisés (Smeele, 1994). Le gain audiovisuel sera donc modulé par la saillance visuelle des phonèmes à identifier (van Wassenhove et al., 2005 ; Arnal et al., 2009). Les consonnes articulées à l'avant du conduit vocal devraient bénéficier d'un large bénéfice audiovisuel (Smeele, 1994) par rapport à la condition auditive, mais également d'un fort impact de la modalité visuelle, surtout au début de la séquence (Jesse & Massaro, 2010) alors que les places d'articulation plus en arrière devraient amener une augmentation graduelle du taux d'identification. Nous faisons l'hypothèse qu'un gain audiovisuel sera observé quand les deux modalités sont complémentaires, celui-ci se manifestant par une identification plus précoce lors de la présentation audiovisuelle ainsi que par des performances de détection plus importantes lors de la présentation bimodale. Cependant, le bénéfice audiovisuel devrait être moins important, voire inexistant, quand l'une ou l'autre des modalités contient assez d'informations pour permettre l'identification.

Ce pattern a été la plupart du temps obtenu dans nos données. L'avantage audiovisuel sur les deux informations unimodales n'a été que rarement observé puisque les résultats obtenus dans la présentation audiovisuelle sont souvent portés principalement par l'une ou l'autre des modalités au travers du temps (visuelle pour /p/, auditive puis visuelle pour /t/). Cela signifie qu'aussi bien d'un point de vue temporel qu'au niveau des scores, la présentation audiovisuelle ne permet pas de détection plus précoce et plus importante que les deux autres modalités. En effet, quand les participants perçoivent audiovisuellement des phonèmes dont les caractéristiques articulatoires ou acoustiques sont saillantes, l'addition des informations des autres modalités (qui sont moins informatives), ne permettent pas de bénéfice particulier par rapport à la présentation unimodale, et perturbe même l'utilisation des autres informations. Les indices les plus informatifs ont le plus grand impact sur la décision, guidant une détection basée sur l'information unimodale. Ainsi, on constate que la détection de /p/ est principalement basée sur l'utilisation des indices visuels. La détection des consonnes /k/ et /t/

devrait être quant à elle être guidée principalement par les informations auditives puisque les indices visuels ne sont pas suffisants pour atteindre 100% de DC même lorsque toute la séquence est présentée (les pourcentages de DC se stabilisent en effet autour de 80-85% de DC). Cela est d'ailleurs bien le cas pour /k/ pour lequel le dévoilement du *burst* permet une augmentation synchronisée des performances lors des présentations auditives et audiovisuelles.

La détection de /t/ est guidée par les deux types d'informations au cours du temps. En effet dans un premier temps, ce sont les informations auditives qui guident la détection puisque la présentation audiovisuelle permet autant de DC que la présentation auditive jusqu'à 210 ms de présentation. Par la suite la tendance s'inverse avec des patterns plus similaires entre les courbes audiovisuelle et visuelle. Ce changement peut s'expliquer par le fait que jusqu'au *gate* 190, l'information auditive sur l'identité de la consonne est disponible notamment via les transitions formantiques puis disparaît par la suite. En effet, alors que pour /p/ la fermeture labiale est déjà complétée, celle de /t/ et /k/ ne le sont pas. Après le *gate* 190, les indices acoustiques (qui sont désormais absents) ne permettent plus d'accumuler de l'information sur l'identité de la consonne. Pour distinguer /t/ de /k/, les participants ne peuvent que se baser sur l'information visuelle avant l'apparition de la seconde voyelle. La réalisation articulatoire de /t/, même si elle est moins saillante que /p/ contient des indices sur lesquels les participants peuvent se baser pour répondre. En effet, la production de la consonne /t/ entraîne une visibilité des dents qui se rejoignent. Cet indice est assez discriminant pour pouvoir détecter /t/ (et donc le distinguer de /k/) puisque le seuil de détection de 90% est dépassé 70 ms plus tôt lors de la présentation audiovisuelle. Rappelons cependant que dans le cas de /t/, si les informations visuelles sont nécessaires pour améliorer la détection avant le *burst*, elles ne sont pas suffisantes puisque la courbe visuelle n'atteint que ponctuellement 90% de DC, de manière tardive et n'atteint jamais 100% de DC. Enfin la détection de /k/ est d'avantage basée sur les informations auditives puisque les seuils de détection (50 et 90%) sont toujours atteints de manière plus tardive lors de la présentation visuelle par rapport aux deux autres modalités de présentation. C'est donc bien la composante auditive qui guide la détection de /k/.

5.5.3 AVANTAGE TEMPOREL DE DETECTION EN FONCTION DE LA SAILLANCE

Si l'on considère uniquement l'impact de l'ajout des informations visuelles, nous obtenons le pattern généralement observé dans la littérature avec un avantage temporel de détection (i.e., la différence temporelle entre le seuil de 90% de détection lors de la présentation audiovisuelle par rapport au gate contenant le *burst*) d'autant plus important que la consonne était articulée à l'avant du conduit vocal (i.e., 160 ms pour /p/, 70 ms pour /t/ et 0 ms pour /k/). Le pattern obtenu était le suivant :[(seuil /p/ AV << seuil /p/ A) << (seuil /t/ AV < seuil /t/ A) < (seuil /k/ AV = seuil /k/ A)]. La distribution temporelle des informations articulatoires permet donc aux participants de détecter les informations concernant la place d'articulation avant la fin de l'intervalle silencieux qui précède le *burst*.

Ces résultats sont en accord avec ceux de Jesse & Massaro (2010) et Smeele (1994) puisque nous obtenons une gradation de l'avantage en fonction de la saillance visuelle des consonnes. Smeele (1994) utilisait un protocole de *gating* mettant en jeu des séquences CV présentées en modalité auditive, audiovisuelle et visuelle. Elle a mis en évidence que les bilabiales et labiodentales étaient identifiées de manière plus précoce quand le signal visuel était disponible. Dans la même veine, Jesse & Massaro (2010) ont mis en évidence que les consonnes articulées à l'avant du conduit vocal bénéficiaient d'avantage de l'ajout des informations visuelles. Ils ont montré qu'en terme de caractéristiques articulatoires, une consonne bilabiale (e.g., /ba/) contient trois fois plus d'information qu'une consonne sans fermeture complète (e.g., /da/), elle est donc plus saillante. Nous obtenons un pattern cohérent avec cette observation. En effet, la présentation visuelle de /p/ bénéficie d'un indice que les deux autres consonnes n'ont pas : la fermeture labiale. Grâce à cet indice, la consonne /p/ est détectée dans 90% des cas seulement 20 ms après la fermeture labiale (*gate* 190), soit 160 ms avant le début le *burst*. Alors que la fermeture labiale est suivie de silence, voir cette fermeture est suffisant pour une détection robuste de la consonne (Smeele, 1994). On observe en effet une similarité très importante entre les patterns obtenus lors des présentations audiovisuelle et visuelle indiquant que les informations articulatoires sont dans ce cas particulièrement informatives et permettent de prédire l'identité de la consonne bien avant le relâchement du *burst* acoustique. Comme ce fût le cas dans l'étude de Jesse & Massaro (2010), les indices visuels sont utilisés très rapidement après leur apparition puisqu'ils permettent de détecter la consonnes dans un cas sur deux (seuil de 50%) seulement 50 ms après le début de la fermeture labiale (*gate* 70). Une augmentation des taux de détection de 50 à 90% se fait alors lors de la présentation audiovisuelle en l'espace de cinq *gates* (soit 50 ms). Il n'en faudra que trois pour que les scores augmentent de la même manière lors de la présentation visuelle seule.

Cette utilisation rapide des indices n'est pas observée dans le cas des consonnes dont la place d'articulation est plus en arrière. Celles-ci produisent une augmentation graduelle du taux de détection, les scores passant de 50 à 90% de DC en 16 et 24 *gates* pour les consonnes /t/ et /k/ respectivement lors de la présentation audiovisuelle. La consonne /t/ bénéficie des informations visuelles dans une moindre mesure que /p/ durant la période de 210 à 360 ms. La reconnaissance se base dans un premier temps principalement sur l'information auditive pour ces deux consonnes, ce qui est généralement observé pour discriminer des places d'articulation situées plus en arrière du conduit vocal (Jesse & Massaro, 2010). Cependant dès que les informations acoustiques sur le *burst* sont disponibles, c'est sur cette information que se basent les participants pour répondre. Enfin, concernant /k/, au départ comme à la fin du signal, les informations auditives tendent à être les plus utilisées.

Les consonnes /t/ et /k/ ont notamment besoin de plus d'information pour être identifiées à hauteur de 90% lors de la présentation audiovisuelle car les indices visuels sont moins saillants que pour /p/. Il faudra en effet attendre les *gates* 310 et 370 pour atteindre le même pourcentage de DC pour /t/ et /k/ respectivement. Dans ce sens, l'avantage temporel de détection observé pour /p/ (l'avance par rapport au dévoilement du *burst*) diminue pour les consonnes moins saillantes avec 70 ms d'avance lors de la présentation audiovisuelle de /t/ par rapport au relâchement du *burst* et une absence d'avantage dans le cas de /k/ (un avantage de 160 ms était obtenu pour /p/). Ces derniers résultats ne sont pas en accords avec ceux de Jesse et Massaro qui obtenaient un avantage audiovisuel pour toutes les consonnes de leur étude, même si celles qui étaient articulées à l'avant du conduit vocal bénéficiaient d'un avantage plus important. Cette différence en terme de résultat peut s'expliquer par la nature de la tâche puisqu'ils utilisaient une tâche de reconnaissance de mots. Or nous savons que la présentation audiovisuelle favorise l'activation lexicale (Fort, 2011). Il est donc probable que les consonnes dont l'articulation est moins visible bénéficient tout de même d'indices qui peuvent être renforcés via les activations lexicales ce qui n'est pas le cas dans le cadre de la perception de syllabes ou séquences CVC.

5.5.4 DE L'APPARITION DES INFORMATIONS VISUELLES A LEUR UTILISATION

Chandrasekaran et al. (2009) ainsi que Schwartz & Savariaux (2013, 2014) ont observé le déroulement temporel de l'apparition des informations visuelles et auditives dans le cadre de la production de la parole. Alors que Chandrasekaran et ces collaborateurs ont mis en

évidence une apparition des informations visuelles bien avant l'information auditive, les données de Schwartz & Savariaux (2013, 2014) ont permis de tempérer ces résultats. Dans leur étude, ils précisent que l'anticipation de l'information visuelle sur l'information auditive lors de la production de la parole survenait uniquement dans des cas très spécifiques, tels que des situations de syllabes CV isolées ou bien en début d'énoncé. Selon les chercheurs, dans de telles situations, les gestes produits seraient des gestes préparateurs alors que dans des situations où les syllabes sont incluses dans des séquences, comme nous pouvons l'observer la plupart du temps dans la parole, les gestes produits seraient des « *co-modulatory gestures* » (i.e., gestes co-modulateurs). Ainsi, dans les gestes préparateurs, la source visuelle serait présente avant la source auditive alors que les gestes co-modulateurs, quant à eux, fourniraient des informations auditives et visuelles de manière plus ou moins synchrone. Leur expérience en production utilisait des séquences CVs ou C pouvait correspondre à /b, d, g/, /p, t, k/, /m, n/ et où V correspondait à la voyelle /a/. La séquence CV était produite soit de manière isolée, soit incluse dans une séquence de type VCVCVCVC. Ces séquences étaient prononcées par un locuteur français. Les résultats ont montré, pour les syllabes isolées, une large anticipation de plus de 150 ms de la source visuelle sur la source auditive. Pour ce qui est des syllabes incluses dans une séquence, les auteurs ont observé que les informations auditives et visuelles étaient plus ou moins synchrones, l'anticipation de l'information visuelle était très inférieure à 150 ms. Selon ces auteurs, il y aurait des cas où l'information visuelle serait disponible avant l'information auditive (e.g., CV), d'autres où les sources visuelle et auditive seraient synchrones (e.g., /aCa/) et enfin d'autres cas où l'information auditive anticiperait l'information visuelle (e.g., lorsque la visibilité des gestes articulatoires est pauvre). Ces résultats, tout comme ceux de Chandrasekaran et al. (2009) se basent sur des données de production.

Pour notre part, et malgré l'utilisation de séquences aCa, nous observons d'une part que les informations visuelles sont présentes avant le signal acoustique et ce à hauteur de 300 ms, du début de la fermeture labiale au *burst*. Bien sur cela ne vaut que pour un seul exemplaire de chaque séquence. L'observation du set original d'enregistrements nous indique que cette durée était en moyenne de 313 ms (SD = 10 ms) pour /p/, 320 ms (SD = 10 ms) pour /t/ et enfin de 316 (SD = 9 ms) pour /k/ en contexte vocalique /a/ (moyennes effectuées sur trois exemplaires). Alors que ces informations peuvent être considérées comme le début de l'information visuelle exploitable, nous constatons à travers nos données que celles-ci ne sont pas systématiquement exploitées dès leur apparition. En effet, le début de la fermeture ne

contient pas d'information permettant de détecter une consonne. Ces mouvements anticipatoires sont « neutres » car les caractéristiques spécifiques à l'articulation d'une consonne ne sont disponibles que plus tard. Par exemple, afin de discriminer /p/ de /t/ et /k/, il faut attendre le moment où la fermeture sera plus marquée que lors de la production d'un /k/ ou d'un /t/. Si la fermeture dépasse ce seuil, il est probable que la consonne perçue soit une bilabiale. Pour distinguer /k/ de /t/, il faudra encore attendre l'apparition d'autres indices spécifiques de production de l'une ou l'autre. L'apparition plus marquée des dents du locuteur orientera les choix vers une consonne dentale plus qu'alvéolaire qui laisserait plutôt entrevoir la langue à l'arrière du conduit vocal, mais qui ne sera pas marquée ni au niveau labial, ni dental. Dans notre cas, il faut par exemple attendre 90 ms après le début de la fermeture (soit 100 ms avant la fermeture) pour que les informations visuelles soient suffisamment discriminantes pour détecter /p/ à hauteur de 50% sur la base des informations visuelles (présentée de manière unimodale ou de concert avec les informations auditives) alors que la fermeture complète des lèvres permettra instantanément d'identifier la consonne. Tout comme pour /p/, /t/ bénéficiera également d'un indice qui permettra, lors de la présentation visuelle, d'atteindre le seuil de 50% de manière précoce, au *gate* 160. En effet, à partir de 130 ms, alors que la mandibule remonte, la langue est mobilisée afin de rejoindre le point de contact (i.e., les dents) qui permettra la production du /t/, cela s'accompagnant d'une légère fermeture buccale. Cet indice, disponible sur une période de 130 à 180 ms permet d'atteindre le seuil de 50% de DC. Nous pouvons noter que le dépassement de ce seuil est plus tardif lors de la présentation visuelle qu'audiovisuelle ou auditive (respectivement 160, 150, 130). Cela est en grande partie dû au fait que les participants se basent plus, au début du signal, sur l'information acoustique pour répondre. En effet, ce ne sera qu'à partir du *gate* 210, que la détection de /t/ commencera à se baser sur les informations visuelles, puisque les scores en audiovisuels seront, jusqu'au *gate* 370, principalement dûs à l'information visuelle.

5.5.5 DESAVANTAGE MULTIMODAL POUR LA DETECTION

On constate un léger désavantage en terme temporel pour atteindre le seuil de 90% de DC lors de la présentation audiovisuelle de /p/ par rapport à la présentation visuelle. Celle-ci est en effet plus précoce lors de la présentation visuelle (*gate* 190) que lors de la présentation audiovisuelle (*gate* 210). Cet avantage temporel lors de la présentation visuelle n'est cependant pas associé à une augmentation des performances puisque les scores visuels sont équivalents à ceux de la modalité audiovisuelle (qui permet 78% de DC pour ce *gate*). Ces

deux observations nous indiquent que la fermeture complète du conduit vocal est un indice déterminant pour l'identification de cette consonne mais que celui-ci n'est pas utilisé de la même façon lors de la présentation audiovisuelle. En effet, lors de la présentation bimodale, il faut attendre 20 ms supplémentaires pour que les participants dépassent ce seuil. Il s'avère qu'au niveau du *gate* 190, où la fermeture est totale, l'ajout de l'information auditive amène de l'ambiguïté qui ne s'explique que par la présence des informations auditives. En effet, lors de la présentation des informations visuelles seule, un seul indice visuel (la fermeture labiale) est utilisable, lors de la présentation bimodale, deux informations peuvent être prises en compte. Or, l'information auditive est à ce moment plus ambiguë car elle ne donne accès qu'à des indices moins marqués sur l'identité de la consonne (transition formantique). Il est possible que l'association des deux modalités entraîne ce délai de 20 ms car celui-ci serait nécessaire pour analyser à la fois les informations visuelles (la fermeture labiale comme un indice spécifique à la plosive /p/) ainsi que les indices auditifs (les transitions formantiques). Cela reflète peut être le délai nécessaire à l'intégration. Cet effet n'est cependant que transitoire puisqu'au *gate* 210, déjà, les informations visuelle et audiovisuelle permettent de nouveau des pourcentages de DC équivalents entre eux, mais supérieurs à ceux obtenus lors de la présentation auditive. Il semble donc que l'ajout de l'information auditive aux informations visuelles soit délétère pour la détection mais seulement au *gate* où l'information visuelle de fermeture est disponible.

5.5.6 QUESTIONNEMENTS METHODOLOGIQUES

Il est à noter que les résultats précédemment décrits souffrent du fait qu'un seul exemplaire de chaque séquence a été utilisé. Ce nombre restreint d'exemplaire était lié, d'un part, à la nouveauté du matériel utilisé (notamment la caméra rapide) qui a induit des erreurs d'enregistrement, rendant inutilisable la moitié des enregistrement et d'autre part, à la correspondance temporelle entre les séquences qui devait être à 20 ms prêt. De plus, rappelons que ce point a également rendu l'analyse statistique des données plus complexes puisqu'aucun point d'isolation n'a pu être déterminé à partir des données individuelles. L'expérience durant une heure, il n'était pas non plus envisageable de doubler (voire tripler) le nombre de listes (afin de fournir plusieurs données individuelles sur un *gate* donné). Même si des précautions ont été mises en place pour palier aux problèmes liés à l'utilisation d'un exemplaire unique (i.e., fade-in visuel, rejets des participants ayant un comportement qui

atteste de l'utilisation d'indices non pertinents, etc), il reste envisageable que l'utilisation de plusieurs exemplaires modulent les résultats obtenus. Dans ce sens, un corpus d'items a été réenregistré afin de reconduire cette étude.

5.6 CONCLUSION GENERALE

Même si un avantage audiovisuel similaire à celui obtenu précédemment dans la littérature n'a pas été observé (Jesse & Massaro, 2010 ; Smeele, 1994), les résultats ont tout de même apporté des éléments nouveaux.

La résolution temporelle fine utilisée à permis de mettre en évidence des variations rapides induites par certains indices. Par exemple, seul 50 ms sont nécessaires pour que les performances augmentent de 50 à 90% lors de la présentation audiovisuelle de /apa/, et les mêmes seuils sont dépassés en 30 ms lors de la présentation visuelle.

De plus, le protocole a également permis de mettre en évidence, en plus de l'évolution progressive du taux de réponse déjà observé par Jesse & Massaro (2010) ou Smeele (1994), des plateaux (autour de 65 % de DC) dans le processus de reconnaissance sur la base des informations auditives pour les trois consonnes. Durant ces instants (du *gate* 170 à 370), l'identification est mise en attente, dans notre cas, sur une période d'environ 200 ms.

En accord avec les résultats de Jesse & Massaro (2010), le signal visuel fourni des informations sur la place d'articulation. Ces informations sont celles qui sont préférentiellement utilisées pour les places d'articulation situées à l'avant du conduit vocal (e.g., fermeture labiale totale est caractéristique des bilabiales comme /p/). Les consonnes articulées plus à l'arrière bénéficient d'indices visuels moins discriminants qui entraînent des variations dans l'utilisation de l'information visuelle. Dans notre cas, la détection de /t/ est basée sur l'information visuelle 100 ms avant le relâchement du *burst*, car la visibilité des dents est un indice encore suffisamment saillant pour qu'il soit exploité. Ce n'est pas le cas pour /k/, dont l'absence de caractéristique visuellement distinctive entraîne une reconnaissance principalement basée sur l'information auditive.

5.7 RESUME

But de l'étude : Examiner l'évolution de l'identification de phonèmes au travers du temps pour un panel de trois consonnes du français qui varient en terme de la saillance perceptive. Celle-ci à également pour but de valider un protocole expérimental (i.e., *gating on-line*).

Populations : 35 participants francophones ont participé à l'étude. Neufs ont été retirés des analyses.

Protocole : Le protocole de cette expérience avait deux particularités. La première était qu'un protocole de Gating à été adapté de façon a ce que les participants réponde durant les présentation (et non après). La seconde était que les *gates* présentés duraient 10ms offrant ainsi nue résolution temporelle importante. Pour cette première expérience, les stimuli testés était les consonnes plosives /p t k/ insérées dans des séquences de type /aCa/. Au début d'une liste, le participants été informé de la consonne cible qu'il devait détecté, puis chaque *gate* des trois séquences lui était présenté de façon aléatoire. La tâche du participant était de répondre aussi vite que possible des qu'il pensait percevoir la consonne cible. L'opération été répétée neuf fois afin que toutes les consonnes aient été détecté dans chacune des modalités de présentation (auditive, audiovisuelle et visuelle).

Résultats : Les données ont montré un effet de la saillance avec un dépassement du seuil de reconnaissance (90%) 160 ms avant le *burst* lorsque les mouvements articulatoires du /p/ été disponible, 70 ms avant le *burst* lorsque l'articulation était celle de /t/ et une absence de bénéfice pour /k/. Dans certaines conditions, il semble également que la présentation bimodale ne permette pas systématiquement un avantage sur les présentations unimodales. Le seuil de détection de 90% est par exemple atteint plus tardivement lors de la présentation audiovisuelle comparé à la présentation visuelle seule.

Conclusion : Nous avons observé que l'apport de la modalité visuelle n'est pas systématique et que la prédictibilité de l'identité du phonème dépend de la saillance visuelle des mouvements articulatoires du locuteur.

CHAPITRE 7

DISCUSSION GENERALE

LIMITES ET PERSPECTIVES

6.1 RAPPEL DES PRINCIPAUX RESULTATS

L'objectif de ce travail était d'étudier l'apport des informations visuelles fournies par les mouvements labiaux lors de la perception de phonèmes non natifs. Nous avons dans un premier temps examiné si des monolingues étaient capables, grâce aux informations labiales, de surmonter le phénomène de surdité phonologique (Etude 1). Nous avons pour cela testé deux populations de monolingues francophones et hispanophones sur leur capacité à distinguer des contrastes phonologiques qui n'existent pas dans leur langue (/f/-/θ/ et /b/-/v/ respectivement) en présentation auditive seule et audiovisuelle. Nous avons observé qu'en présentation auditive seule, les francophones étaient phonologiquement sourds au contraste /f/-/θ/ de l'espagnol. En revanche, en présentation audiovisuelle, ils étaient capables d'utiliser les informations visuelles fournies par les gestes faciaux du locuteur pour discriminer les deux phonèmes. De plus, l'accès à ces informations permettait d'accélérer les temps de réponses, comme c'est le cas lors de la perception de phonèmes natifs. En d'autres termes, l'information sur la gestualité oro-faciale a permis aux participants francophones de surmonter la surdité phonologique dont ils étaient victimes en situation de présentation auditive, et cela en l'absence de connaissance du visème présenté (i.e., de l'interdentale fricative /θ/). Pour les hispanophones, nous n'avons pas observé de surdité phonologique pour le contraste /b/-/v/ en présentation auditive seule.

Nous avons par la suite exploré, auprès de participants francophones, les processus neuronaux précoces qui se mettent en place lors de la perception audiovisuelle d'un phonème non natif (Etude 2). Le même paradigme, ainsi que le même contraste phonologique (i.e., /f/-/θ/, qui n'existe pas en français) ont été utilisés. Notre attention portait principalement sur le complexe N1/P2 des potentiels évoqués auditifs. La présentation audiovisuelle produit une réduction quasi systématique de l'amplitude de N1 auprès des participants francophones ainsi qu'auprès du groupe de contrôle d'hispanophones. Cela atteste de l'allègement des traitements auditifs lorsque des informations visuelles sur la gestualité oro-faciale sont présentées. Ceci est en accord avec des études antérieures comme celle de van Wassenhove et al. (2005), mais la modulation de l'amplitude ne s'accompagne pas d'une réduction de la latence. Les modulations observées sur P2 dans la littérature n'ont pas été obstenues. Elles variaient en fonction du phonème présenté et de la modalité de présentation. Ces différences peuvent s'expliquer par des différences acoustiques et visuelles entre les stimuli. Dans notre expérience nous avons utilisé des consonnes fricatives, alors que les études antérieures ont toujours porté sur des plosives. Les indices auditifs présents dans le signal de friction sont

plus distribués que ceux de la plosion car le temps de relâchement de la friction est plus important ; celle-ci atteint donc son amplitude maximum plus lentement que les plosives (Johnson, 2003 ; Kluender & Walsh, 1992 ; Walsh & Diehl, 1991). Concernant leurs attributs visuels, même si les mouvements articulatoires étaient présents à partir du même moment pour toutes les séquences, il n'est pas exclu que leurs propriétés (comme la saillance) aient modulé les résultats observés. La contribution la plus intéressante de ce travail est que nous avons observé, chez les francophones, une composante très précoce de type P50 qui apparaît autour de 30 ms lorsqu'on traite visuellement des phonèmes qui n'existent pas dans notre langue. Ce type de réponse P50 est généralement associée, lors de la présentation auditive, à des stimuli pour lesquels nous ne sommes pas « habitués » (Grunwald et al., 2003). Dans notre cas, celle-ci émergeait lors de la présentation audiovisuelle de mouvements labiaux non natifs. Malgré le faible nombre de participants du groupe contrôle d'hispanophones, les résultats mettent en lumière un processus spécifique, pré-attentif, lors des processus perceptifs de phonèmes non natifs. Celui-ci module sans doute les traitements auditifs subséquents. D'autres études seront nécessaires afin de déterminer son rôle avec plus de précision.

Par la suite, nous nous sommes interrogés sur l'utilisation des informations sur les mouvements labiaux par des individus ayant bénéficiés d'un contact précoce avec plusieurs langues (Etude 3). La question était de savoir si une expérience précoce avec un double code phonologique avait un impact sur la manière d'exploiter les informations visuelles sur la gestualité oro-faciale lors du traitement des phonèmes non natifs. Nous avons mis en place une tâche de discrimination auditive et audiovisuelle dans laquelle des participants bilingues et monolingues devaient identifier un phonème qui n'existe pas dans leur répertoire phonologique : le phonème dental rétroflexe /t̪/. Ce phonème est souvent confondu avec le phonème /t/ qui existe dans les langues parlées par les participants. Les résultats ont montré que les performances ainsi que les temps de réponse des monolingues et les bilingues ne différaient pas lors de la présentation auditive : ils étaient tous sourds phonologiquement au contraste /t̪/-/t/. Il semblerait donc que face à des phonèmes inconnus, les bilingues ne diffèrent pas des monolingues. Les deux populations parvenaient néanmoins à améliorer leur performance et surmonter la surdité phonologique lors de la présentation audiovisuelle, attestant une fois de plus qu'il n'est pas nécessaire de connaître le visème pour pouvoir exploiter des informations visuelles sur son articulation. Cependant, lors de la présentation audiovisuelle, les résultats ont révélé que les bilingues tirent moins partie des informations visuelles puisque leur taux de réponses correctes était inférieur à celui des monolingues. De plus les bilingues étaient également plus lents pour répondre lors de la présentation

audiovisuelle. Nous avons supposé la mise en jeu de processus de traitement de visages différents entre ces deux populations, qui seraient plus coûteux pour les bilingues en terme de traitement et qui serait moins efficace pour identifier les phonèmes non natifs.

Enfin l'étude 4 présente les résultats préliminaires d'une étude qui avait pour but de quantifier de manière fine les apports de chaque modalité dans le processus de détection de consonnes natives. Pour ce faire nous avons mis en place un protocole de *gating* (pas de 10 ms) avec des présentations auditive, visuelle et audiovisuelle des phonèmes consonantiques /p/, /t/ et /k/ insérés dans des séquences VCV. Les traitements ont révélé d'une part qu'un bénéfice audiovisuel n'est pas systématiquement observable. Alors qu'au seuil de 50%, peu de différences sont observées entre les modalités de présentation, des différences ont émergé pour le seuil de détection de 90%. Nous avons alors observé un effet de la modalité ainsi que de la saillance visuelle des consonnes, avec des détections d'autant plus précoces que l'information était saillante : (seuil /p/ AV << seuil /p/ A) << (seuil /t/ AV < seuil /t/ A) < (seuil /k/ AV = seuil /k/ A). Ce bénéfice allait de 160 ms pour /p/ à 0 ms pour /k/. De plus, nous avons également observé que les participants se basaient sur des informations différentes en fonction de la consonne. Alors que /k/ est principalement reconnu sur la base des informations auditives, les deux autres consonnes sont reconnues sur la base de différentes informations au cours du temps. Avant la fermeture labiale de /p/, il semble que les informations auditives soient les plus utilisées par les participants, malgré le fait que celles-ci ne contiennent pas assez d'information pour permettre de dépasser le seuil de 60% de détection correcte. Après la fermeture labiale, la reconnaissance se basait sur les informations visuelles qui, même présentées seules, permettent d'atteindre très rapidement 90% de réponses correctes. Après l'apparition du *burst*, n'importe quelle information peut être utilisée afin de favoriser la détection. Concernant le phonème /t/, les informations auditives semblent être celles qui permettent le plus de reconnaissances correctes jusqu'à la période de silence avant le *burst*. A ce moment, même si les participants détectent la consonne à hauteur de 60%, les informations visuelles seront utilisées préférentiellement, sans doute car celles-ci permettent de distinguer /t/ de /k/.

6.2 PERSPECTIVES ET LIMITES

6.2.1 APPORT DES INFORMATIONS VISUELLES LORS DE LA PERCEPTION DES PHONEMES NON NATIFS : DU COMPORTEMENT A LA NEUROPHYSIOLOGIE

Dans le cadre de la perception des phonèmes natifs, une facilitation liée à la présentation audiovisuelle a été largement observée dans la littérature (lors de la perception dans le bruit ou sans bruit (Erber, 1969 ; MacLeod & Summerfield, 1987 ; Ross et al., 2007 ; Sumby & Pollack 1954 ; Grant & Seitz, 2000 cité par Schwartz et al., 2004 ; Cerrato, Leoni et Falcone, 1998 ; (Christian Benoît et al., 1994). Certains auteurs pensent qu'elle serait due au fait que les informations fournies par les mouvements articulatoires permettent de prédire le contenu acoustique qui va suivre. Cependant, le débat concernant la nature de cette facilitation reste encore ouvert. Certaines études mettent en évidence que les *indices de forme* permettent un effet d'amorçage qui réduit le coût des traitements subséquents (notamment l'activité du cortex auditif comparé à une condition sans indice) en permettant de prédire quelles caractéristiques acoustiques vont suivre (Jääskeläinen et al., 2008; Paris et al., 2013) et d'autres, que des *indices temporels* permettent une amélioration des scores d'identification (Schwartz et al., 2004).

Outre la nature de ces prédictions, une question reste encore sans réponse. Est-ce que les phonèmes inconnus pourraient bénéficier des mêmes prédictions lorsque ceux-ci sont inconnus? Nous savons que pour des participants ayant des connaissances préalables de la langue étrangère, un bénéfice est généralement observé à partir du moment où le phonème impliqué est assez visible. Quelques études impliquant des individus monolingues avaient déjà permis de montrer que des séquences non natives étaient mieux perçues et répétées lors d'une présentation bimodale (Davis & Kim, 1998, 2001). Notre travail nous permet de fournir une bribe de réponse supplémentaire. Les études 1 et 3 ont permis de mettre en évidence des réponses plus précoces lors de la présentation audiovisuelle de phonèmes non natifs. Tout comme pour les phonèmes natifs (Paris et al., 2013 ; Reisberg, McLean, & Goldfield, 1987), le processus d'identification est plus efficace et accéléré lors de la présentation audiovisuelle.
Notons cependant que le bénéfice audiovisuel n'est observable que si il y a un phénomène de surdité phonologique lors de la présentation auditive seule (ce qui n'était pas le cas des hispanophones pour le contraste /b/-/v/ dans l'Etude 1), et qu'avec des phonèmes qui seraient suffisamment *saillants* pour que l'informations visuelles puissent être utilisées par un locuteur

étranger (Dodd, 1977 ; Hazan et al., 2006 ; Sennema, Hazan, & Faulkner, 2003 ; Wang, Behne, & Jiang, 2009 ; voir Hardison, 2003 pour des résultats montrant une amélioration pour le contraste /r/-/l/ peu visible). En effet, de nombreux facteurs peuvent impacter l'utilisation de ces indices. Cependant et malgré la richesse des résultats déjà obtenus depuis les 15 dernières années sur la perception audiovisuelle de la parole non native, le domaine souffre d'un manque d'unicité, et notamment de modèle fédérateur qui guiderait la démarche expérimentale.

L'investigation des processus neurophysiologiques mis en place lors de la perception de phonèmes non natifs (Etude 2) a permis de montrer que lorsque ces phonèmes sont présentés en modalité audiovisuelle, ceux-ci entraînent également une modulation de l'amplitude de la composante auditive précoce N1, comme c'est le cas pour les phonèmes natifs (Besle et al., 2004; van Wassenhove et al., 2005). Nous avons conjointement observé l'apparition d'une P50 lors de la perception d'un phonème non natif par des francophones, celle-ci étant d'ailleurs la seule condition où une réduction de l'amplitude de la P2 a été observée. Cette P50 serait donc possiblement une réponse du système face à des mouvements préparatoires inconnus ou des entrées « considérées » comme incongruentes. Afin d'aller plus loin dans les interprétations, nous devons poursuivre les passations avec des participants hispanophones supplémentaires.

6.2.2 LA PERCEPTION AUDIOVISUELLE DES PHONEMES NON NATIFS

La littérature concernant la perception audiovisuelle de la langue maternelle est riche et a permis de générer des modèles de perception audiovisuelle de la parole. Des modèles de perception auditive des phonèmes non natifs existent (e.g., « *Perceptual Assimilation Model* » de Best, 1995 ou le « *Speech Learning Model* » de Flege, 1995), mais rares sont les auteurs ayant modélisés son versant visuel. Le modèle de Hazan et al. (2006) est le seule à notre connaissance à formuler des hypothèses concernant les rapports entres les visèmes de la L1 et de la L2. Rappelons que dans ce modèle, trois catégories de visèmes sont distinguables et modulent la facilité avec laquelle un individu pourra les utiliser lors de la perception de contrastes non natifs. La première catégorie visuelle concerne les visèmes qui existent à la fois dans la L1 et la L2 et qui marquent les mêmes distinctions phonémiques (e.g., /f-s/ en français et espagnol). L'utilisation des indices visuels se fera alors sans difficulté puisque les deux visèmes sont connus et utilisés dans la langue maternelle. La deuxième catégorie,

contredit par nos résultats, renferme les visèmes qui existent dans la L2 mais pas dans la L1 (e.g., /θ/ espagnol (Etude 1) pour les français et /ʈ/ bengali (Etude 3) pour les français et les bilingues). Dans cette catégorie, les individus sont sensés, selon le modèle, ne pas pouvoir exploiter l'information visuelle. Nous avons constaté que les résultats des études 1 et 3 contrastent avec les prédictions de ce modèle. En effet, malgré le fait que nous ayons utilisé des phonèmes dont la réalisation articulatoire *n'existe pas* dans le répertoire visèmique des participants (ce qui est le cas de l'interdentale espagnole et du phonème rétroflexe du bengali), ceux-ci se sont révélés capables d'utiliser l'information visuelle sur l'identité des phonèmes. De plus, l'information visuelle présente lors de l'articulation d'un phonème inconnu a également fait émerger une P50 au niveau du cortex auditif ce qui atteste du fait que ces informations, lorsqu'elles sont inconnues, modulent les traitements auditifs (Etude 2). Remarquons cependant qu'outre l'existence ou non d'un visème dans la L1, de nombreux facteurs impactent l'utilisation des informations visuelles sur les gestes de production du locuteur. Les caractéristiques visuelles et acoustiques de la langue maternelle (Sekiyama & Tohkura, 1993; Wang et al., 2009) ainsi que la saillance des visèmes étudiés (Dodd, 1977; Hazan et al., 2006) ou la quantité d'expérience dans la L2 (Wang et al., 2008) vont également moduler l'apport de l'information visuelle dans le processus d'identification des phonèmes.

La plupart des études menées sur la perception audiovisuelle des phonèmes non natifs portent sur des populations d'apprenants, généralement de niveau intermédiaire/élevé et/ou résidant dans un pays étrangers (Hardison, 2003; Hazan et al., 2006; Ortega-Llebaria et al., 2001; Wang et al., 2008, 2009; Janet F Werker et al., 1992). Or les résultats obtenus dans ces études ne nous permettent pas d'inférer de manière directe quel sera le comportement d'individus sans connaissance préalable dans une langue donnée face à un visème qu'ils ne connaissent pas. En effet, même s'il semblerait logique qu'un continuum soit dessiné entre les populations naïves et les apprenants, il n'en est rien. Les apprenants ont déjà mis en place un processus de catégorisation des phonèmes étrangers qui est la plupart du temps erroné à cause du phénomène d'assimilation alors que les débutants n'ont aucune connaissance préalable concernant le système phonologique de la L2. Il est possible que pour les apprenants, ce processus de catégorisation des nouveaux phonèmes étant déjà enclenché, les informations visuelles ne soient pas suffisantes pour limiter ces erreurs de catégorisation, surtout si l'on considère que plus on a de connaissances dans une langue, moins on se base sur les informations visuelles. Celles-ci n'auraient donc plus un poids suffisant pour pouvoir améliorer la perception (les participants se basant d'avantage sur l'information auditive). Cela

pourrait expliquer pourquoi nos participants font bon usage des informations visuelles alors que des apprenants échouent à utiliser ces indices lorsqu'ils ne sont pas partagés dans la L1 et la L2. Cela souligne également l'importance de sensibiliser les individus aux caractéristiques articulatoires de la langue étrangère dès le début de l'apprentissage. Cependant, afin de statuer sur la validité de cette hypothèse, d'autres investigations seront nécessaires, notamment en comparant directement des groupes de participants dont le niveau de maîtrise de la L2 se situerait du niveau « débutant » à « élevé ».

Une étude de Werker et al. (1992) fournit quelques pistes sur l'impact de l'information visuelle sur les gestes du locuteur en fonction du niveau de maîtrise de la L2. Les auteurs se sont dans un premier temps attachés à montrer une surdité phonologique lors de la présentation auditive de phonèmes, notamment de l'interdentale fricative à des participants québécois francophones. Les individus ont alors reporté beaucoup d'erreurs de place d'articulation. Dans un second temps, ils ont réalisé une tâche de type McGurk dans laquelle un /ba/ auditif était couplé à l'articulation de /ba, va, ða, da, ʒa, ga/ afin de montrer que la perception de la parole visuelle est également impactée par l'expérience linguistique. La tâche contrôle consistait en la présentation des mêmes séquences en modalité visuelle seule ou auditive seule. Les résultats ont montré que le pourcentage de capture visuelle /ða/ était moins important lorsque le niveau de maîtrise de la langue était faible. Les participants répondaient majoritairement /ta/ ou /da/ au lieu de l'interdentale, attestant d'une difficulté à percevoir des visèmes qui n'existent pas dans la langue maternelle, mais également d'un phénomène « d'assimilation visuelle ». Cela va à l'encontre de nos résultats puisque même lors de la présentation visuelle seule, les participants parvenait à identifier l'interdentale fricative. Il est donc judicieux de questionner l'impacte des méthodologies employées, notamment dans la formulation des réponses, qui étaient orales dans l'étude de Werker et al. (1992), celles-ci étant enregistrées à l'aide d'un dictaphone puis retranscrites par des individus natifs ayant des notions en phonétique.

Signalons enfin que même si les Etudes 1, 2 et 3 permettent de tirer des conclusions intéressantes, qui divergent parfois des résultats obtenus dans la littérature, celles-ci sont difficiles à inscrire dans une vision unifiée de la perception audiovisuelle de la parole non native. En effet, celle-ci est impactée par de nombreux facteurs et relève de prédictions spécifiques en fonction de (1) la structure du répertoire phonologique natif, (2) le poids donnée aux informations visuelles dans la langue maternelle, (3) l'existence ou non des visèmes considéré dans la langue maternelle, ainsi que (4) leur saillance visuelle et (5) l'expérience de l'individu avec la L2.

6.2.3 AMELIORATION DU PROCESSUS D'IDENTIFICATION SANS ENTRAINEMENT

Au travers de la littérature, nous pouvons constater que l'entraînement auditif est souvent utilisé pour pallier aux problèmes induits par la surdité phonologique. Néanmoins, ceux basés sur les informations auditives ne sont pas très efficaces (10% en général avec une amélioration maximum de 18% observée par Iverson, Hazan et Bannister, 2005), sont longs à mettre en place (2 à 4 semaines d'entraînement) et ne permettent souvent que peu de généralisation à de nouveaux items (Kraus et al., 1995; Lively et al., 1994; Logan et al., 1991). Les entraînements audiovisuels, même s'ils sont peu étudiés, semblent dans certains cas plus efficaces que les entraînements auditifs sur des populations d'apprenants (Hardison, 1999, 2003) alors que certains l'évaluent comme aussi efficace que l'entraînement auditif (Hazan et al., 2005). Cependant, rappelons que, là encore, les populations étudiées sont déjà familiarisées avec la langue puisque les participants testés dans ces études avaient un niveau « intermédiaire » en anglais. Cela ne permet pas, encore une fois, de déterminer quel serait l'impact de tel entraînement sur des populations de monolingues naïfs. De plus, comme le signalaient Heeren & Schouten (2008), les améliorations de performance observées suite à des entraînements en laboratoire ne peuvent être considérées comme des preuves d'apprentissage de représentations phonémiques. Nos résultats, portant sur des monolingues naïfs indiquent que les informations visuelles pourraient être utilisées de manière automatique et même *sans entraînement* dans le cadre de la discrimination de syllabes mais également moduler de façon automatique les activations du cortex auditif. Il semble donc que la simple présentation audiovisuelle puisse se substituer, dans certains cas, aux entraînements auditifs, voire audiovisuels, en offrant le bénéfice d'avoir un impact immédiat sur les processus d'identification phonémique. Enfin, signalons que nos participants n'avaient pas pour but d'acquérir de connaissances phonologiques sur les langues testés, nous ne pouvons aller plus en avant dans nos conclusions. Il serait donc intéressant de pousser plus loin les investigations, notamment via des études longitudinales afin d'étudier l'impact de la présentation audiovisuelle sur la mise en place de nouvelles catégories phonologiques.

6.2.4 L'IMPORTANCE DE L'ETUDE DES POPULATIONS SANS CONNAISSANCE PREALABLE

Alors que la littérature s'est particulièrement intéressée aux apprenants, il est cependant cruciale, selon nous, de savoir si des individus seraient, dès le début de l'apprentissage, capables d'utiliser les informations visuelles fournies par les mouvements articulatoires visibles qui n'existent pas dans la L1. Leur utilisation permettrait à une personne débutant son apprentissage d'une langue étrangère de mieux comprendre les différences articulatoires entre les phonèmes de sa langue maternelle et ceux de celle qu'il apprend. Ainsi deux sons qui sont assimilés lors de la présentation auditive pourraient, sur la base des différences articulatoires visibles, être discriminés et placés dans une nouvelle catégorie du répertoire phonologique et ainsi faciliter l'apprentissage. Nos résultats montrent, dans le cadre d'expérimentation en laboratoire, que la présentation audiovisuelle permet une large amélioration de la perception des phonèmes non natifs.

6.2.5 L'INFORMATION VISUELLE DANS LA PERCEPTION DES PHONEME NATIFS

L'identification des phonèmes est capitale dans la perception de la parole, aussi bien dans notre langue maternelle que lors de l'apprentissage d'une langue étrangère. Elle est la pierre angulaire de la reconnaissance des mots. Si un seul trait phonétique est mal perçu ou mal interprété, c'est toute la suite du processus de traitement de la parole, notamment l'analyse phonologique et sémantique, qui sera mis en échec. Prenons comme exemple le mot « cagoulé ». Alors que nous sommes spécialisés dans la perception et l'identification des sons de notre langue, nous ne pourrons sans doute jamais faire une analyse erronée de ce mot dans des conditions d'écoutes adéquates. Mais dans des conditions d'écoutes adverses, par exemple lors d'une conversation téléphonique à l'extérieur, il est tout à fait envisageable d'entendre « taboulé » juste par une confusion de place d'articulation (i.e., /k/ et /t/ qui sont des consonnes plosives non voisées et /g/ et /b/ qui sont des consonnes plosives voisées). En effet, des études ont montré que la place d'articulation, contrairement au mode d'articulation, est souvent peu audible mais très visible (Jesse & Massaro, 2010; Smeele & Sitting, 1991). Dans des cas comme celui-ci, il a été montré que l'apport des informations visuelles est important, même lors de l'identification des phonèmes de notre langue maternelle. L'étude 4, à permis de mettre en évidence une augmentation importante et rapide des scores de détection lors de la présentation audiovisuelle par rapport à la présentation auditive (e.g., /p/ est détecté dans 90%

des cas 20 ms après la fermeture labiale et une augmentation de 50 à 90% de détection correctes est observée en l'espace de 50 ms). De plus, le seuil de détection est plus précoce lors de la présentation audiovisuelle. En effet, alors qu'il faut de manière systématique attendre le *burst* pour que les informations auditives permettent le dépassement du seuil de 90%, l'ajout des informations visuelles permet de détecter /p/ 160 ms plus tôt et /t/ 60 ms avant plus tôt.

6.2.5.1 LIMITE

La limitation majeure de cette étude est liée à l'utilisation d'un seul exemplaire par séquence. En effet, le design expérimental ainsi que les contraintes imposées par le *gating*, notamment sur la sélection des stimuli, n'ont pas permis d'utiliser plusieurs exemplaires de chaque séquence ce qui limite le potentiel de généralisation des résultats. Malgré cette limitation, les résultats ont permis de montrer un effet de la saillance visuelle ainsi qu'une détection globalement plus précoce lors de la présentation audiovisuelle comme précédemment obtenu dans la littérature (Cathiard, 1994; Jesse & Massaro, 2010; Paris et al., 2013; Smeele, 1994). Un nouveau corpus a été enregistré afin de reconduire cette expérience dans de meilleures conditions.

6.2.5.2 PERSPECTIVES

Si les résultats qui seront obtenus avec le nouveau corpus sont encourageants et confirment la tendance des présents résultats, nous envisageons de reconduire cette expérience avec des phonèmes non natifs afin de comprendre comment nous utilisons les indices auditifs et visuels lorsque nous surmontons le phénomène de surdité phonologique en situation de présentation audiovisuelle (Wang et al., 2008; Werker et al., 1992). La présentation des phonèmes non natifs en situation auditive, visuelle et audiovisuelle nous permettra, grâce à la tâche du *gating*, de « décortiquer » le processus d'intégration bimodale ainsi que de mieux comprendre comment et dans quelle mesure nous tirons partie de chacun des canaux.

RÉFÉRENCES BIBLIOGRAPHIQUES

Abel, J., Barbosa, A., Black, A., Mayer, C., & Vatikiotis-Bateson, E. (2011). The labial viseme reconsidered: Evidence from production and perception. *The Journal of the Acoustical Society of America*, *129*(4), 2456. doi:10.1121/1.3588075

Abramson, A. S., & Lisker, L. (1970). The voicing Dimension: Some experiments in comparative phonetics. In *International Congress of Phonetic Sciences* (pp. 563–567). Prague.

Abramson, A. S., & Lisker, L. (1973). Voice-timing perception in Spanish word-initial stops. *Journal of Phonetics*, *1*, 1–8.

Abry, C., & Boë, L.-J. (1980). *Labialité et Phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiales*. Publications de l'Université des langues et lettres de Grenoble.

Abry, C., & Lallouache, M.-T. (1995). Le MEM : un modèle d'anticipation paramétrable par locuteur : Données sur l'arrondissement en français. *Les Cahiers de l'ICP. Bulletin de La Communication Parlée*, (3), 85–99.

Abry, C., Lallouache, M.-T., & Cathiard, M.-A. (1996). How can coarticulation models account for speech sensitivity to audio visual desynchronization? In D. Stork & M. Hennecke (Eds.), *Speechreading by Humans and Machines* (pp. 247–255). Berlin: Springer-Verlag.

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A Systematic Review and Meta-Analysis of the Cognitive Correlates of Bilingualism. *Review of Educational Research*, *80*, 207–245. doi:10.3102/0034654310368803

Aloufy, S., Lapidot, M., & Myslobodskym, M. (1996). Differences in susceptibility to the "blending illusion" among Native Hebrew and English speakers. *Brain and Language*, *53*, 51–57. doi:10.1006/brln.1996.0036

Altarriba, J., & Heredia, R. R. (2008). *An introduction to bilingualism: Principles and processes*. New York: Lawrence Erlbaum Associates.

Andersson, U., Lyxell, B., Rönnberg, J., & Spens, K.-E. (2001). Cognitive Predictors of Visual Speech Understanding. *Journal Deaf Studies and Deaf Education*, *6*, 103–115.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics*, *32*(2), 233–250. doi:10.1016/S0095-4470(03)00036-6

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience*, *29*(43), 13445–53. doi:10.1523/JNEUROSCI.3194-09.2009

Arnold, P., & Hill, F. (2001). Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*, 339–355. doi:10.1348/000712601162220

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, *8*, 170–177. doi:10.1016/j.tics.2004.02.010

Auer, E. T. (2002). The influence of the lexicon on speech read word recognition: contrasting segmental and lexical distinctiveness. *Psychonomic Bulletin & Review*, *9*, 341–347. doi:10.3758/BF03196291

Auer Jr, E. T., & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *The Journal of the Acoustical Society of America*, *102*, 3704–3710. doi:10.1121/1.420402

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*, 115–21. doi:10.1016/j.neuropsychologia.2013.11.011

Baart, M., & Vroomen, J. (2010). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neuroscience Letters*, *471*(2), 100–3. doi:10.1016/j.neulet.2010.01.019

Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition*, *130*(1), 31–43. doi:10.1016/j.cognition.2013.09.006

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you "read" tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, *52*(6), 493–503. doi:10.1016/j.specom.2010.03.002

Barrós-Loscertales, A., Ventura-Campos, N., Visser, M., Alsius, A., Pallier, C., Avila Rivera, C., & Soto-Faraco, S. (2013). Neural correlates of audiovisual speech processing in a second language. *Brain and Language*, *126*(3), 253–62. doi:10.1016/j.bandl.2013.05.009

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*, 809–823. doi:10.1016/S0896-6273(04)00070-4

Benguerel, A.-P., & Pichora-Fuller, M. K. (1982). Coarticulation Effects in Lipreading. *Journal of Speech and Hearing Research*, *25*(4), 600–607.

Benoît, C., Guiard-Marigny, T., Le Goff, B., & Adjoudani, A. (1996). Which components of the face do humans and machines best speechread? *Speechreading by Humans and Machines*, *150*, 315–328.

Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of Phonetic Context on Audio-Visual Intelligibility of French. *Journal of Speech and Hearing Research*, *37*, 1195–1203.

Berger, M. (1951). *The American English pronunciation of Russian immigrants*. Columbia University, New York.

Bernstein, L. E., Demorest, M. E., Coulter, D. C., & O'Connell, M. P. (1991). *Lipreading sentences with vibrotactile vocoders: performance of normal-hearing and hearing-impaired subjects. The Journal of the Acoustical Society of America* (Vol. 90, pp. 2971–2984). doi:10.1121/1.401771

Bernstein, L. E., Iverson, P., & Auer Jr, E. T. (1997). Elucidating the complex relationships between phonetic perception and word recognition in audiovisual speech perception. In *Proceedings of the ESCA/ESCOP Workshop on Audio-VisualSpeech Processing* (pp. 89–92). Rhodes, Greece.

Berthommier, F. (2004). A Phonetically Neutral Model of the Low-level Audiovisual Interaction. *Speech Communication*, *44*, 31–41.

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *The European Journal of Neuroscience*, *20*(8), 2225–34. doi:10.1111/j.1460-9568.2004.03670.x

Best, C. C., & McRoberts, G. W. (2003). Infant Perception of Non-Native Consonant Contrasts that Adults Assimilate in Different Ways. *Language and Speech*, *46*, 183–216.

Best, C. T. (1991). The emergence of Native-Language Phonological influences in Infants : A perceptual assimilation model. *Haskins Laboratories Status Report on Speech Research*, 1–30.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In *The development of speech perception: The transition from speech sounds to spoken words* (Vol. 167, pp. 167–224).

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research,* (StrangeW.,., pp. 167–200.).

Best, C. T., & Avery, R. A. (1999). Left-Hemisphere Advantage for Click Consonants is Determined by Linguistic Significance and Experience. *Psychological Science*. doi:10.1111/1467-9280.00108

Best, C. T., Goodell, E., Womer, J., Insabella, G., Klatt, L., Luke, S., & Silver, J. (1990). Infant and adult perception of nonnative speech contrasts differing in relation to the listeners' native phonology. In *International Conference in Infant Studies*. Montreal.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America*, *109*, 775–794.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology*, *14*, 45–60.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988a). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology*, *14*, 345–360. doi:10.1037/0096-1523.14.3.345

Best, C. T., & Tyler, M. (2007a). Nonnative and second-language speech perception: Commonalities and complementarities. In M. BohnO-S. (Ed.), *Language Experience in Second language Speech Learning* (In honor o., pp. 13–34).

Best, C. T., & Tyler, M. D. (2007b). Nonnative and Second-Language speech perception: Commonalities and Complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Second Language Speech Learning: the Role of Language Experience in Speech Perception and Production* (pp. 13–34). Amsterdam : John Benjamins.

Bhat, J., Pitt, M. a., & Shahin, A. J. (2014). Visual context due to speech-reading suppresses the auditory response to acoustic interruptions in speech. *Frontiers in Neuroscience*, *8*, 1–9. doi:10.3389/fnins.2014.00173

Bialystok, E. (2008). Bilingualism: The good, the bad, and the indifferent. *Bilingualism: Language and Cognition*, *12*(1), 3. doi:10.1017/S1366728908003477

Bialystok, E., Craik, F., & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(4), 859–73. doi:10.1037/0278-7393.34.4.859

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research*, *17*, 619–630. doi:10.3758/bf03211678

Blair, R. C., & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology*, *30*, 518–524. doi:10.1111/j.1469-8986.1993.tb02075.x

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer [version 5.1.42].

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and Ultimate attainment in the pronunciation of foreign Language. *Studies in Second Language Acquisition*, *19*, 447–465.

Boomershine, A., Hall, K. C., Hume, E., & Johnson, K. (2008). The impact of allophony versus contrast on speech perception. In P. Avery, E. B. Dresher, & K. Rice (Eds.), *Contrasts in Phonology: Theory, Perception, Acquisition* (pp. 143–172). Berlin & New York : Mouton de Gruyter.

Boulenger, V. (2006). *Le Langage et l'Action : Dynamique des liens fonctionnels unissant verbes d'action et contrôle moteur*. Université Lumière Lyon II, France.

Boutros, N. N., Barkerb, B. A., Tuetingb, P. A., Wub, S., & Nasrallahb, H. A. (1995). The P50 evoked potential component and mismatch detection in normal volunteers : implications for the study of sensory gating. *Psychiatry Research*, *57*, 83–88.

Boutros, N. N., & Belger, A. (1999). Midlatency evoked potentials attenuation and augmentation reflect different aspects of sensory gating. *Biological Psychiatry*, *45*(7), 917–922. doi:10.1016/S0006-3223(98)00253-4

Boysson-Bardies, B. (1996). *Comment la parole vient aux enfants*. Paris: Odile Jacob.

Bradlow, A. R. (1993). *Language-specific and universal aspects of vowel production and perception : a cross linguistic study of vowel inventories*. Cornell University, Ithaca, Etats-Unis.

Brannen, K. (2002). The role of perception in differential substitution. *Canadian Journal of Linguistics*, *47*, 1–46.

Braun, B., Lemhöfer, K., & Mani, N. (2011). Perceiving unstressed vowels in foreign-accented English. *The Journal of the Acoustical Society of America*, *129*, 376–387. doi:10.1121/1.3500688

Brown, C. (2000). The interrelation between speech perception and phonological acquisition from infant to adult. In J. Archibald (Ed.), *Second Language Acquisition and Linguistic Theory*. Oxford : Blackwell.

Brown, C. A. (1998). The role of the L1 grammar in the L2 acquisition of segmental structure. *Second Language Research*, *14*, 136–193. doi:10.1191/026765898669508401

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327. doi:10.1111/j.2044-8295.1986.tb02199.x

Burnham, D. (2003). Language specific speech perception and the onset of reading. *Reading and Writing*, *16*, 573–609. doi:10.1023/A:1025593911070

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–20. doi:10.1002/dev.20032

Burnham, D. H. (1998). Language specificity in the development of auditory-visual speech perception. In D. Burnham, R. Campbell, & B. Dodd (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory visual speech* (pp. 27–60). Hove, UK: Lawrence Erlbaum Associates Ltd.

Burns, T. C., Werker, J. F., & Mcvie, K. (2003). Development of Phonetic Categories in Infants Raised in Bilingual and Monolingual Environments. In B. et al. Beachley (Ed.), *Proceedings of the 27th annual Boston University Conference on Lnaguage Development* (pp. 173–184). Boston MA: Cascadilla Press.

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2003). Neural correlates of cross-modal binding. *Nature Neuroscience*, *6*(2), 190–5. doi:10.1038/nn993

Byers-Heinlein, K., Burns, T. C., & Werker, J. F. (2010). The roots of bilingualism in newborns. *Psychological Science*, *21*(3), 343–8. doi:10.1177/0956797609360758

Calbour, C., & Dumont, A. (2002). *Voir la parole : lecture labiale, perception audiovisuelle de la parole.* Paris, Masson.

Callan, D. E., Jones, J. a, Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *NeuroImage*, *22*(3), 1182–94. doi:10.1016/j.neuroimage.2004.03.006

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, *14*, 2213–2218. doi:10.1097/00001756-200312020-00016

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., & Akahane-Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *NeuroImage*, *19*(1), 113–124. doi:10.1016/S1053-8119(03)00020-X

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex*, *11*, 1110–1123. doi:10.1093/cercor/11.12.1110

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport*, *10*, 2619–2623. doi:10.1097/00001756-199908200-00033

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Steven, C. R., Mcguire, P. K., … Williams, S. C. R. (1997). Activation of Auditory Cortex During Silent Lipreading. *Science*, *276*(5312), 593–596.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649–657. doi:10.1016/S0960-9822(00)00513-3

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 1001–10. doi:10.1098/rstb.2007.2155

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., … David, A. S. (2001). Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research*, *12*, 233–243. doi:10.1016/S0926-6410(01)00054-4

Cathiard, M.-A. (1994). *La perception visuelle de l'anticipation des gestes vocaliques : cohérence des évènements audibles et visibles dans le flux de la parole*. Université Stendhal, Grenoble, France.

Cathiard, M.-A., & Tibergihien, G. (1994). Le visage de la parole : Une cohérence bi-modale temporelle ou configurationelle ? *Psychologie Française*, *39*, 357–374.

Cerrato, L., Leoni, F. A., & Falcone, M. (1998). Is it Possible to Evaluate the Contribution of Visual Information to the Process of Speech Comprehension ? In *AVSP 98, International Conference on Auditory-Visual Speech Processing* (pp. 141–146). Terrigal, Australia.

Chandrasekaran, C., & Ghazanfar, A. a. (2009). Different neural frequency bands integrate faces and voices differently in the superior temporal sulcus. *Journal of Neurophysiology*, *101*(2), 773–88. doi:10.1152/jn.90843.2008

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS One Computational Biology*, *5*(7), e1000436. doi:10.1371/journal.pcbi.1000436

Chen, T. H., & Massaro, D. W. (2004). Mandarin speech perception by ear and eye follows a universal principle. *Perception & Psychophysics*, *66*, 820–836. doi:10.3758/BF03194976

Cheour, M., Shestakova, A., Alku, P., Ceponiene, R., & Näätänen, R. (2002). Mismatch negativity shows that 3-6-year-old children can learn to discriminate non-native speech sounds within two months. *Neuroscience Letters*, *325*(3), 187–90.

Coles, M. G. H., & Rugg, M. D. (1995). Event-related brain potentials: an introduction. In M. D. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of Mind: Event-related Brain Potentials and Cognition* (pp. 1–26). New York, NY, US: Oxford University Press. doi:10.1093/acprof:oso/9780198524168.003.0001

Colin, C., & Radeau, M. (2003). Les illusions McGurk dans la parole : 25 ans de recherches. *L'année Psychologique*, *103*(3), 497–542. doi:10.3406/psy.2003.29649

Colin, C., Radeau, M., Soquet, A., Dachy, B., & Deltenre, P. (2002). Electrophysiology of spatial scene analysis: The mismatch negativity (MMN) is sensitive to the ventriloquism illusion. *Clinical Neurophysiology*, *113*, 507–518. doi:10.1016/S1388-2457(02)00028-7

Conboy, B. T., & Mills, D. L. (2006). Two languages, one developing brain: Event-related potentials to words in bilingual toddlers. *Developmental Science*, *9*(1), 1–12. doi:10.1111/j.1467-7687.2005.00453.x

Conboy, B. T., Sommerville, Jessica, A., & Kuhl, P. K. (2008). Cognitive control factors in speech perception at 11 months. *Developmental Psychology*, *44*(5), 1505–12. doi:10.1037/a0012975

Cook, V., Bassetti, B., Kasai, C., Sasaki, M., & Takahashi, J. a. (2006). Do bilinguals have different concepts? The case of shape and material in Japanese L2 users of English. *International Journal of Bilingualism*, *10*(2), 137–152. doi:10.1177/13670069060100020201

Cook, V. J. (1997). The consequences of bilingualism for cognitive processing. In M. B. de Groot & J. F. Kroll (Eds.), *Tutorials in Bilingualism: Psycholinguistic Perspectives* (pp. 279–299). Mahwah, N.J.: Lawrence Erlbaum Associates.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: implications for models of lexical access. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*, 1283–1296. doi:10.1037/0278-7393.26.5.1283

Costa, A., Hernández, M., & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, *106*, 59–86. doi:10.1016/j.cognition.2006.12.013

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics*, *5*, 365–373. doi:10.3758/BF03210660

Crystal, D. (2003). *English as a Global Language*. (p. 212). Cambridge: Cambridge University Press.

Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication*, *52*, 555–564. doi:10.1016/j.specom.2010.02.006

Davis, C., & Kim, J. (1998). Repeating and Remembering Foreign Language Words: Does seeing help ? In *AVSP 98, International Conference on Auditory-Visual Speech Processing* (pp. 6–10). Terrigal - Sindey, Australia.

Davis, C., & Kim, J. (1999). Perception of Clearly Presented Foreign Language Sounds: The Effects of Visible Speech. In *AVSP 99, International Conference on Auditory-Visual Speech Processing* (pp. 1–6). Santa Cruz, CA, USA.

Davis, C., & Kim, J. (2001). Repeating and remembering foreign language words: Implications for language teaching systems. *Artificial Intelligence Review*, *16*, 37–47. doi:10.1023/A:1011086120667

Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. *The Quarterly Journal of Experimental Psychology*, *57*, 1103–1121. doi:10.1080/02724980343000701

De la Vaux, S. K., & Massaro, D. W. (2004). Audiovisual speech gating: examining information and information processing. *Cognitive Processing*, *5*(2), 106–112. doi:10.1007/s10339-004-0014-2

De Saussure, F. (1916). *Cours de linguistique générale*. *Collection Payothèque* (p. 509).

Dehaene-lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception. *NeuroReport*, *8*(4), 919–924.

Dehaene-Lambertz, G., & Gliga, T. (2004). Common neural basis for phoneme processing in infants and adults. *Journal of Cognitive Neuroscience*, *16*, 1375–1387. doi:10.1162/0898929042304714

Dekeyser, R. M. (2000). The Robustness of Critical Period Effects in Second Language Acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.

Delattre, P. (1955). Acoustic Loci and Transitional Cues for Consonants. *The Journal of the Acoustical Society of America*, *52*, 127–137. doi:10.1121/1.1917767

Demorest, M. E., & Bernstein, L. E. (1992). Sources of variability in speechreading sentences: a generalizability analysis. *Journal of Speech and Hearing Research*, *35*, 876–891.

Descout, R., Boë, L.-J., & Abry, C. (1980). Labialité vocalique et labialité consonantique. Un jeu des lèvres au féminin : l'idiolecte. In *Labialité et Phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiales* (pp. 111–126).

Diehl, R. L., & Kluender, K. R. (1989). On the Objects of Speech Perception. *Ecological Psychology*, *1*, 121–144. doi:10.1207/s15326969eco0102_2

Dodd, B. (1977). The role of vision in the perception of speech. *Perception*, *6*, 31–40. doi:10.1068/p060031

Dodd, B., McIntosh, B., & Woodhouse, L. (1998). Early lipreading ability and speech and language development of hearing-impaired pre-schoolers. In R. Campbell, B. Dodd, & D. K. Burnham (Eds.), *Hearing by Eye (II): The Psychology of Speechreading and Auditory-visual Speech* (pp. 229–242). Hove, UK: Lawrence Erlbaum Associates Ltd.

Dornyei, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition. The Psychology of the Language Learner: Individual Differences in Second Language Acquisition.*

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *The Journal of the Acoustical Society of America*, *97*, 585–592. doi:10.1121/1.413112

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, *95*, 2670–2680. doi:10.1121/1.409836

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: a perceptual illusion ? *Journal of Experimental Psychology*, *25*(6), 1568–1578.

Dupoux, E., Sebastián-Gallés, N., Pallier, C., & Mehler, J. (1997). A Destressing " Deafness " in French ? *Journal of Memory*, *36*, 406–421.

Eggermont, J. J., & Ponton, C. W. (2002). The neurophysiology of auditory perception: From single units to evoked potentials. *Audiology and Neuro-Otology*, *7*, 71–99. doi:10.1159/000057656

Eimas, P. D. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the [r-l] distinction by young infants. *Perception & Psychophysics*, *18*(5), 341–347. doi:10.3758/BF03211210

Eimas, P. D., Siqueland, Ei. R., Jusczyk, P. W., & Vigorito, J. (1971). Speech perception in Infants. *Science*, *171*(3968), 303–306.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*, 423–425.

Erdener, D. V, & Burnham, D. (2013). The relationship between auditory-visual speech perception and language-specific speech perception at the onset of reading instruction in English-speaking children. *Journal of Experimental Child Psychology*, *116*(2), 120–38. doi:10.1016/j.jecp.2013.03.003

Erdener, D. V, & Burnham, D. K. (2005). The Role of Audiovisual Speech and Orthographic Information in Nonnative Speech Production. *Language Learning*, *55*(June), 191–228.

Escudero, P. (2005). *Linguistic Perception and Second Language Acquisition : Explaning the attainment of Optimal Phonological categorization*. Utrecht University, Pays-Bas.

Escudero, P., & Boersma, P. (2004). Bridging the Gap Between L2 Speech Perception Research and Phonological Theory. *Studies in Second Language Acquisition, 26*, 551–585. doi:10.1017/S0272263104040021

Escudero, P., & Polka, L. (2003). A Cross-language Study of Vowel categorization and Vowel Acoustics. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 861–864). Barcelona, Spain.

Escudier, P., Benoît, T., & Lallouache, C. (1990). Identification visuelle de stimuli associés à l'opposition /i/-/y/ : Etude statique. *Journal de Physique Colloques*, *51*, 541–544.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles : a TMS study. *European Journal of Neuroscience*, *15*(2002), 399–402.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, *11*, 796–804.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language. Evidence for the effect of equivalence classification. *Journal of Phonetics*, *15*, 47–65.

Flege, J. E. (1995). Second Language Speech Learning: Theory, Findings, and Problems. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233–277). Timonium, MD: York Press. doi:10.1111/j.1600-0404.1995.tb01710.x

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470. doi:10.1006/jpho.1997.0052

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5), 3125–34.

Flege, J. E., & Port, R. (1981). Cross-Language Phonetic Interference: Arabic to English. *Language and Speech*, *24*, 125–146. doi:10.1177/002383098102400202

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age Constraints on Second-Language Acquisition. *Journal of Memory and Language*, *41*(1), 78–104. doi:10.1006/jmla.1999.2638

Fort, A., Delpuech, C., Pernier, J., & Giard, M.-H. (2002). Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans. *Cerebral Cortex*, *12*, 1031–1039. doi:10.1093/cercor/12.10.1031

Fort, M. (2011). *L'accès au lexique dans la perception audiovisuelle et visuelle de la parole*. Université de grenoble, france.

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2012). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes*. doi:10.1080/01690965.2012.701758

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*, 3–28. Retrieved from http://129.237.66.221/P800/Fowler1986.pdf

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*, 1730–1741. doi:10.1121/1.415237

Fowler, C. A. (2004). Speech as a Supramodal or Amodal Phenomenon. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Mulisensory Processes* (pp. 189–201). Cambridge, MA: MIT Press.

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, *68*, 161–177. doi:10.3758/BF03193666

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Experimental Psychology*, *17*, 816–828. doi:10.1037/0096-1523.17.3.816

Fox, R. A., Flege, J. E., & Munro, M. J. (1995). The perception of English and Spanish vowels by native English and Spanish listeners: a multidimensional scaling analysis. *The Journal of the Acoustical Society of America*, *97*, 2540–2551. doi:10.1121/1.411974

Freedman, R., Adler, L. E., Waldo, M. C., Pachtman, E., & Franks, R. D. (1983). Neurophysiological evidence for a defect in inhibitory pathways in schizophrenia: comparison of medicated and drug-free patients. *Biological Psychiatry*, *18*, 537–551.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The Motor Theory of Speech Perception Reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361–377.

Gentil, M. (1981). Etude de la perception de la parole : Lecture labiale et sosies labiaux. *Technical Report,* IBM, France.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*, 473–490. doi:10.1162/089892999563544

Gibson, E. J. (1969). *Principles of perceptual learning and development. Principles of perceptual learning and development.* (p. 353). New York: Appleton-Century-Crofts.

Giraud, A.-L., & Poeppel, D. (2012). Speech Perception from a Neurophysiological Perspective. In D. Poeppel, T. Overath, A. Popper, & R. Fay (Eds.), *The Human Auditory Cortex* (pp. 225–245). New York: Springer.

Golestani, N., Molko, N., Dehaene, S., LeBihan, D., & Pallier, C. (2007). Brain structure predicts the learning of foreign speech sounds. *Cerebral Cortex*, *17*(3), 575–82. doi:10.1093/cercor/bhk001

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: Reallocation of neural substrates. *NeuroImage*, *21*, 494–506. doi:10.1016/j.neuroimage.2003.09.071

Gollan, T. H., & Acenas, L.-A. R. (2004). What is a TOT? Cognate and translation effects on tip-of-the-tongue states in Spanish-English and tagalog-English bilinguals. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(1), 246–69. doi:10.1037/0278-7393.30.1.246

Grant, K. W. (2003). Auditory Supplements to Speechreading. In *Proceedings of the Institute for Electronics, Information and Communication Engineers (IEICE) and The Acoustical Society of Japan, Speech Dynamics by Ear, Eye, Mouth and Machine: An Interdisciplinary Workshop* (pp. 1–5). Kyoto, Japan.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*, 1197–1208. doi:10.1121/1.422512

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *The Journal of the Acoustical Society of America*, *100*, 2415–2424. doi:10.1121/1.417950

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*, 2677–2690. doi:10.1121/1.422788

Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*. doi:10.1037/0012-1649.25.4.577

Grosjean, F. (1980). Spoken word recognition processes. *Perception & Psychophysics*, *28*(4), 267–283.

Grosjean, F. (1996). Gating. *Language and Cognitive Processes*, *11*(6), 597–603.

Grunwald, T., Boutros, N. N., Pezer, N., Von Oertzen, J., Fernández, G., Schaller, C., & Elger, C. E. (2003). Neuronal substrates of sensory gating within the human brain. *Biological Psychiatry*, *53*, 511–519. doi:10.1016/S0006-3223(02)01673-6

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America*, *100*, 1111–1121. doi:10.1121/1.416296

Guiard-Marigny, T., Tsingos, N., Adjoudani, A., Benoit, C., & Gascuel, M.-P. (1996). 3D models of the lips for realistic speech animation. *Proceedings Computer Animation '96*. doi:10.1109/CA.1996.540490

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: the case of Japanese adults' perception of English

consonants. *The Journal of the Acoustical Society of America*, *107*, 2711–2724. doi:10.1121/1.428657

Hamann, S., & Sennema, A. (2005). Voiced labiodental fricatives or glides - all the same to Germans? In *PSP 2005 ISCA Workshop on Plasticity in Speech Perception*. London, UK.

Hannon, E. E., & Trehub, S. E. (2005). Metrical categories in infancy and adulthood. *Psychological Science*, *16*(1), 48–55. doi:10.1111/j.0956-7976.2005.00779.x

Hardison, D. M. (1999). Bimodal Speech Perception by Native and Nonnative Speakers ofEnglish-Factors Influencing the Mcgurk Effect. *Language Learning*, *41*(6), 3–73.

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*(04), 495–522. doi:10.1017/S0142716403000250

Hardison, D. M. (2005). Variability in bimodal spoken language processing by native and nonnative speakers of English: A closer look at effects of speech style. *Speech Communication*, *46*(1), 73–93. doi:10.1016/j.specom.2005.02.002

Harnsberger, J. D. (2001). The perception of Malayalam nasal consonants by Marathi, Punjabi, Tamil, Oriya, Bengali, and American English listeners: A multidimensional scaling analysis. *Journal of Phonetics*, *29*(3), 303–327. doi:10.1006/jpho.2001.0140

Hausmann, M., Durmusoglu, G., Yazgan, Y., & Güntürkün, O. (2004). Evidence for reduced hemispheric aysmmetries in non-verbal functions in bilinguals. *Journal of Neurolinguistics*, *17*, 285–299. doi:10.1016/S0911-6044(03)00049-6

Hayashi, Y., & Sekiyama, K. (1998). Native-Foreign Language Effect in the McGur Effect: a Test With Chinese and Japanese. In *AVSP 98, International Conference on Auditory-Visual Speech Processing* (pp. 61–66). Terrigal, Sydney, Australia.

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics*, *28*, 377–396. doi:10.1006/jpho.2000.0121

Hazan, V., Sennema, A., & Faulkner, A. (2002). Audiovisual Perception in L2 Learners. In *Proceedings of the International Conference for Spoken Language Processing* (pp. 1685–1688). London, UK.

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America*, *119*(3), 1740. doi:10.1121/1.2166611

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, *47*(3), 360–378. doi:10.1016/j.specom.2005.04.007

Hebb, D. O. (1949). The organization of behavior: a neuropsychological theory. *Science Education*, *44*, 335. doi:10.2307/1418888

Heeren, W. F. L., & Schouten, M. E. H. (2008). Perceptual development of phoneme contrasts: how sensitivity changes along acoustic dimensions that contrast phoneme categories. *The Journal of the Acoustical Society of America*, *124*, 2291–2302. doi:10.1121/1.2967472

Heeren, W. F. L., & Schouten, M. E. H. (2010). Perceptual development of the Finnish /t-t:/ distinction in Dutch 12-year-old children: A training study. *Journal of Phonetics*, *38*, 594–603. doi:10.1016/j.wocn.2010.08.005

Heider, F., & Heider, G. (1940). An experimental investigation of lipreading. *Psychological Monographs*, *54*, 124–153.

Hickok, G. (2009). Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans. *Journal of Cognitive Neuroscience*, *21*(7), 1229–1243. doi:10.1162/jocn.2009.21189.Eight

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*, 131–138. doi:10740277

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*, 393–402. doi:10.1038/nrn2113

Hojen, A. D. (2003). *Second-language speech perception and production in adult learners before and after short-term immersion*. University of Aarhus, Danemark.

Hume, E., & Johnson, K. (2003). The Impact of Partial Phonological Contrast on Speech Perception. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1–8). Barcelona, Spain.

Hygge, S., Rönnberg, J., Larsby, B., & Arlinger, S. (1992). Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. *Journal of Speech and Hearing Research*, *35*, 208–215. doi:10.1044/jshr.3501.208

Istria, M., Nicolas-Jeantoux, C., & Tamboise, J. (1982). *Manuel de lecture labiale (exercices d'entraînement)*. (p. 175). Paris: Mason. doi:10.2307/1777736

Iverson, P., Bernstein, L. E., & Auer Jr, E. T. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Communication*, *26*, 45–63.

Iverson, P., Ekanayake, D., Hamann, S., Sennema, A., & Evans, B. G. (2008). Category and perceptual interference in second-language phoneme learning: an examination of English /w/-/v/ learning by Sinhala, German, and Dutch speakers. *Journal of Experimental Psychology*, *34*(5), 1305–16. doi:10.1037/0096-1523.34.5.1305

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America*, *118*(5), 3267. doi:10.1121/1.2062307

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, *97*, 553–562. doi:10.1121/1.412280

Iverson, P., & Kuhl, P. K. (1996). Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/. *The Journal of the Acoustical Society of America*, *99*, 1130–1140. doi:10.1121/1.415234

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*. doi:10.1016/S0010-0277(02)00198-1

Jääskeläinen, I. P., Koskentalo, K., Balk, M. H., Autti, T., Kauramäki, J., Pomren, C., & Sams, M. (2008). Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing. *The Open Neuroimaging Journal*, *2*, 14–19. doi:10.2174/1874440000802010014

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics*, *72*(1), 209–225. doi:10.3758/APP

Jiang, J. (2003). *Relating Optical Speech to Speech Acoustics and Visual Speech Perception*. *Department of Electrical Engineering*. University of california, Los Angeles.

Johannesen, J. K., Kieffaber, P. D., O'Donnell, B. F., Shekhar, A., Evans, J. D., & Hetrick, W. P. (2005). Contributions of subtype and spectral frequency analyses to the study of P50 ERP amplitude and suppression in schizophrenia. *Schizophrenia Research*, *78*, 269–284. doi:10.1016/j.schres.2005.05.022

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60–99.

Johnson, K. (2003). Chapter 8: Stops and Affricates. In Blackwell (Ed.), *Acoustic and Auditory Phonetics* (2nd ed., pp. 135–148).

Johnson, K., & Babel, M. (2010). On the perceptual basis of distinctive features: Evidence from the perception of fricatives by Dutch and English speakers. *Journal of Phonetics*, *38*(1), 127–136. doi:10.1016/j.wocn.2009.11.001

Jusczyk, P. W., Cutler, a, & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, *64*(3), 675–87.

Jusczyk, P. W., & Luce, P. A. (2002). Speech perception and spoken word recognition: past and present. *Ear and Hearing*, *23*, 2–40.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645. doi:10.1006/jmla.1994.1030

Kazanina, N., Phillips, C., & Idsardi, W. (2006). The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(30), 11381–6. doi:10.1073/pnas.0604821103

Kim, J., & Davis, C. (2004). Investigating the audio–visual speech detection advantage. *Speech Communication*, *44*(1-4), 19–30. doi:10.1016/j.specom.2004.09.008

Kizkin, S., Karlidag, R., Ozcan, C., & Ozisik, H. I. (2006). Reduced P50 auditory sensory gating response in professional musicians. *Brain and Cognition*, *61*, 249–254. doi:10.1016/j.bandc.2006.01.006

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279–312.

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research*, *18*(1), 65–75. doi:10.1016/j.cogbrainres.2003.09.004

Kluender, K. R., Diehl, R. L., & Killen, P. R. (1987). Japanese Quail Can Learn Phonetic Categories. *Science*, *237*(4819), 1195–1197.

Kluender, K. R., & Walsh, M. a. (1992). Amplitude rise time and the perception of the voiceless affricate/fricative distinction. *Perception & Psychophysics*, *51*(4), 328–33.

Kluge, D. C. (2009). *Brazilian EFL Learners' Identification of word-final /m-n/: Native/Nonnative realizations and effect of visual cues.* Universidade Federalde Santa Catarina: Florianópolis.

Kluge, D. C., Reis, M. S., Nobre-Oliveira, D., & Bettoni-Techio, M. (2006). The Use of Visual Cues in the Perception of English Syllable-Final Nasals by Brazilian EFL Learners. In M. A. Watkins, A. S. Rauber, & B. O. Baptista (Eds.), *Recent Research in Second Language Phonetics/Phonology: Perception and Production.* (pp. 141–153). Cambridge Scholars Publishing,.

Kochetov, A. (2004). Perception of place and secondary articulation contrasts in different syllable positions: language-particular and language-independent asymmetries. *Language and Speech*, *47*, 351–382. doi:10.1177/00238309040470040201

Koyama, S., Akahane-Yamada, R., Gunji, A., Kubo, R., Roberts, T. P. L., Yabe, H., & Kakigi, R. (2003). Cortical evidence of the perceptual backward masking effect on /l/ and /r/ sounds from a following vowel in Japanese speakers. *NeuroImage*, *18*, 962–974. doi:10.1016/S1053-8119(03)00037-5

Kraus, N., McGee, T., Carelle, T. D., King, C., Tremblay, K., & Nicol, T. (1995). Central Auditory System Plasticity Assiociated with Speech Discrimination Training. *Journal of Cognitive Neuroscience*, *7*(1), 25–32.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107.

Kuhl, P. K. (2000). Language, mind, and brain: Experience alters perception. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 99–115). Cambridge, MA: MIT Press.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews. Neuroscience*, *5*, 831–843. doi:10.1038/nrn1533

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London*, *363*(1493), 979–1000. doi:10.1098/rstb.2007.2154

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science (New York, N.Y.)*, *218*, 1138–1141. doi:10.1126/science.7146899

Kuhl, P. K., & Miller, J. D. (1975). Speech Perception by The Chinchilla: Voiced-Voiceless Distinction in Alveolar Plosive Consonants. *Science, New Series*, *190*(4209), 69–72.

Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla : Identification functions for synthetic VOT stimuli. *Journal of Acoustical Society of America*, *63*, 905–917.

Kuhl, P. K., Stevens, E. B., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for Natives Language phonetic Perception between 6 and 12 Months. *Developmental Science*, *2*(9), F13–F21.

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(15), 9096–101. doi:10.1073/pnas.1532872100

Kuhl, P. K., Tsuzaki, M., Tohkura, Y., & Meltzoff, A. N. (1994). Human processing of auditory-visual information in speech perception: potential for multimodal human–machine interfaces. In *Proc. Inter- nat. Conf. on Spoken Language Proceedings* (pp. 539–542). Tokyo.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. *Science*, *255*, 606–255.

Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *Journal of Experimental Psychology*, *17*, 829–840. doi:10.1037/0096-1523.17.3.829

Laurienti, P. J., Burdette, J. H., Wallace, M. T., Yen, Y.-F., Field, A. S., & Stein, B. E. (2002). Deactivation of sensory-specific cortex by cross-modal stimuli. *Journal of Cognitive Neuroscience*, *14*, 420–429. doi:10.1162/089892902317361930

Lebib, R., Papo, D., de Bode, S., & Baudonnière, P.-M. (2003). Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. *Neuroscience Letters*, *341*(3), 185–188. doi:10.1016/S0304-3940(03)00131-9

Lee, S. A. S., & Iverson, G. K. (2011). Stop consonant productions of Korean–English bilingual children. *Bilingualism: Language and Cognition*, *15*(02), 275–287. doi:10.1017/S1366728911000083

Lengeris, A. (2009). Perceptual Assimilation and L2 Learning: Evidence from Perception of Southern British English Vowels by Native Speakers of Greek anf Japanese. *Phonetica*, *66*, 169–187. doi:10.1159/000235659

Lenneberg, E. (1967). *Biological Foundations of Language*. New York: John, Willey.

Léon, P. (1992). *Phonétisme et prononciation du français* (p. 272). Paris: Nathan-Fac.

Levy, E. S., & Strange, W. (2008). Perception of French vowels by American English adults with and without French language experience. *Journal of Phonetics*, *36*, 141–157. doi:10.1016/j.wocn.2007.03.001

Lewkowicz, D. J. (2010). 15. Infant perception of audio-visual speech synchrony. *Developmental Psychology*, *46*, 66–77. doi:10.1037/a0015579

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(5), 1431–6. doi:10.1073/pnas.1114783109

Liberman, A. M. (1996). *Speech: A Special Code Learning, Development and Conceptual Change:* (MIT Press.). Cambridge, Massachusetts: A Bradford Book.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the Speech Code. *Psychological Review*, *74*(6), 431–461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Some Cues for the Distinction between Voiced and Voiceless Stops in Initial Position. *Language and Speech*, *1*, 153–167. doi:10.1121/1.1919048

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*. doi:10.1037/h0093673

Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, *65*, 497–516. doi:10.2307/1418032

Liberman, A. M., Harris, K., Eimas, P., Lisker, L., & Bastian, J. (1961). An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language & Speech*, *4*, 175–195. doi:10.1177/002383096100400401

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4075760

Lim, S., & Holt, L. L. (2011). Learning foreign sounds in an alien world: videogame training improves non-native speech categorization. *Cognitive Science*, *35*(7), 1390–405. doi:10.1111/j.1551-6709.2011.01192.x

Lindblom, B. E., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical Society of America*, *42*, 830–843. doi:10.1121/1.1910655

Lipski, S. C., & Mathiak, K. (2007). A magnetoencephalographic study on auditory processing of native and nonnative fricative contrasts in Polish and German listeners. *Neuroscience Letters*, *415*(1), 90–5. doi:10.1016/j.neulet.2007.01.001

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese Listeners to Identify English /r/ and /l/. III Long-term retention of New Phonetic Categories. *Journal of Acoustic Society of America*, *96*(4), 2076–2087.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to Identify English /r/ and /l/: A First report. *Journal of Acoustical Society of America*, *89*(2), 874–886.

Lotto, A. J., & Kluender, R. (1998). General contrast effects in speech perception : Effect of Preceding liquid on Stop Consonant Identification. *Perception & Psychophysics*, *60*(4), 602–619.

Lubker, J., & Gay, T. (1982). Anticipatory labial coarticulation: experimental, biological, and linguistic variables. *The Journal of the Acoustical Society of America*, *71*, 437–448. doi:10.1121/1.387447

Luo, H., Liu, Z., & Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS Biology*, *8*, 25–26. doi:10.1371/journal.pbio.1000445

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*, 131–141. doi:10.3109/03005368709077786

Macmillan, N. A., Goldberg, R. F., & Braida, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *The Journal of the Acoustical Society of America*, *84*, 1262–1280. doi:10.1121/1.396626

Major, R. C. (2001). *Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology* (p. 224). Mahwah, NJ: Lawrence Erlbaum Associates.

Major, R. C. (2007). Identifying a Foreign Accent in an Unfamiliar Language. *Studies in Second Language Acquisition*, *29*(4), 539–556. doi:10.1017/S0272263107070428

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(August), 940–967.

Martin, A., & Peperkamp, S. (2011). Speech Perception and Phonology. In M. van Oostendorp, C. Ewen, E. Hume, & K. Rice (Eds.), *Companion to Phonology* (pp. 1–25). Hoboken, N.J. : Wiley-Blackwell.

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development*, *55*, 1777–1788. doi:10.2307/1129925

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* Cambridge, MA: MIT Press.

Massaro, D. W. (2004). From multisensory integration to talking heads and language learning. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 153–176). Cambridge, MA: MIT Press

Massaro, D. W., & Cohen, M. M. (1983). Categorical or Continuous Speech Perception : A New test. *Speech Communication*, *2*, 15–35.

Massaro, D. W., Cohen, M. M., Gesi, A., & Heredia, R. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics*, *21*, 445–478.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *The Journal of the Acoustical Society of America*, *100*, 1777–1786. doi:10.1121/1.417342

Massaro, D. W., Thompson, L. a, Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, *41*(1), 93–113.

Mattock, K., & Burnham, D. (2006). Chinese and English Infants' Tone Perception: Evidence for Perceptual Reorganization. *Infancy*, *10*(3), 241–265. doi:10.1207/s15327078in1003_3

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition*, *106*, 1367–1381. doi:10.1016/j.cognition.2007.07.002

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*, 122–134. doi:10.1111/j.1467-7687.2007.00653.x

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, 102–111. doi:10.1016/S0010-0277(01)00157-3

McCandliss, B. D., Fiez, J. a, Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective & Behavioral Neuroscience*, *2*(2), 89–108.

McClelland, J. L., Fiez, J. a, & McCandliss, B. D. (2002). Teaching the /r/-/l/ discrimination to Japanese adults: behavioral and neural aspects. *Physiology & Behavior*, *77*(4-5), 657–62.

McGrath. (1985). *An examination of cues for visual and audiovisual speech perception using natural and computer generated faces*. University of Nottingham, Angleterre.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *267*, 746–748. doi:10.1038/264746a0

Meunier, C. (2005). Invariants et variabilité en phonétique. In N. N'Guyen, S. Wauquier-Gravelines, & J. Durand (Eds.), *Phonologie et phonétique: Forme et substance* (pp. 349–374). Paris: Hermès.

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America*, *82*, 2145–2147. doi:10.1121/1.395659

Mielke, J., Baker, A., & Archangeli, D. (2010). Variability and homogeneity in American English /r/ allophony and /s/ retraction. In *Variation, Detail, and Representation. (LabPhon 10)* (pp. 699–719). Berlin: Mouton de Gruyter.

Miki, K., Watanabe, S., & Kakigi, R. (2004). Interaction between auditory and visual stimulus relating to the vowel sounds in the auditory cortex in humans: a magnetoencephalographic study. *International Congress Series*, *1278*, 177–180. doi:10.1016/j.ics.2004.11.015

Miller, J. L. (1994). On the internal structure of phonetic categories: a progress report. *Cognition*, *50*(1-3), 271–85.

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *25*(25), 5884–93. doi:10.1523/JNEUROSCI.0896-05.2005

Mills, A. E. (1987). The development of phonology in the blind child. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading* (pp. 145–163). London: Lawrence Erlbaum Associates.

Mills, D. L., Prat, C., Zangl, R., Stager, C. L., Neville, H. J., & Werker, J. F. (2004). Language experience and the organization of brain activity to phonetically similar words: ERP evidence from 14- and 20-month-olds. *Journal of Cognitive Neuroscience*, *16*, 1452–1464. doi:10.1162/0898929042304697

Mohamadi, T., & Benoît, C. (1992). L'Apport de la vision du locuteur à l'intelligibilité de la parole bruitée. *Bulletin de La Communication Parlée*, *2,* Cahier de l'ICP, INP, Grenoble.

Moradi, S., Lidestam, B., & Rönnberg, J. (2013). Gated audiovisual speech identification in silence vs. noise: effects on time and accuracy. *Frontiers in Psychology*, *4*, 359. doi:10.3389/fpsyg.2013.00359

Moradi, S., Lidestam, B., Saremi, A., & Rönnberg, J. (2014). Gated auditory speech perception: effects of listening conditions and cognitive capacity. *Frontiers in Psychology*, *5*, 531. doi:10.3389/fpsyg.2014.00531

Morrison, G. (2006). *L1 & L2 production and perception of English and Spanish vowels: A statistical modelling approach*. University of Alberta, Edmonton, Alberta, Canada.

Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research*, *13*, 417–425.

Möttönen, R., Schürmann, M., & Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: A magnetoencephalographic study. *Neuroscience Letters*, *363*, 112–115. doi:10.1016/j.neulet.2004.03.076

Mourand-Dornier, L. (1980). *Le rôle de la lecture labiale dans la reconnaissance de la parole*. Université de Franche-Comté, France.

Moyer, A. (2007). Do Language Attitudes Determine Accent? A Study of Bilinguals in the USA. *Journal of Multilingual and Multicultural Development*, *28*(6), 502–518. doi:10.2167/jmmd514.0

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*, 351–362. doi:10.3758/BF03206811

Munhall, K. G., Jones, J. a, Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133–7.

Munhall, K. G., & Jones, J. A. (1998). Articulatory evidence for syllabic structure. *Behavioral and Brain Sciences*. doi:10.1017/S0140525X98391268

Munhall, K. G., & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *The Journal of the Acoustical Society of America*, *104*(1), 530–539.

Munhall, K. G., & Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In R. Campbell, D. Dodd, & D. K. Burnham (Eds.), *Hearing by Eye II: Advances in the psychology of speechreading and audiovisual speech* (pp. 123–139). Hove: Psychology Press.

Myers, E. B. (2014). Emergence of category-level sensitivities in non-native speech sound learning. *Frontiers in Neuroscience*, *8*, 238. doi:10.3389/fnins.2014.00238

Myers, E. B., & Swan, K. (2012). Effects of Category Learning on Neural Sensitivity to Non-native Phonetic Categories. *Journal of Cognitive Neuroscience*, *24*, 1695–1708. doi:10.1162/jocn_a_00243

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, *24*, 375–425. doi:10.1111/j.1469-8986.1987.tb00311.x

Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: evidence from nasal place discrimination. *Developmental Science*, *13*(3), 407–20. doi:10.1111/j.1467-7687.2009.00898.x

Navarra, J., Sebastián-Gallés, N., & Soto-Faraco, S. (2005). The perception of second language sounds in early bilinguals: new evidence from an implicit measure. *Journal of Experimental Psychology*, *31*(5), 912. doi:10.1037/0096-1523.31.5.912

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12. doi:10.1007/s00426-005-0031-5

Nittrouer, S. (2004). The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *The Journal of the Acoustical Society of America*, *115*, 1777–1790. doi:10.1121/1.1651192

Ohala, J. J. (1975). The Temporal regulation of speech. In G. Fant & T. MAA (Eds.), *Auditory Analysis and Perception of Speech* (pp. 431–453). London: Academic Press.

Ojanen, V. (2005). *Neurocognitive mechanisms of audiovisual speech perception*. Helsinki University of technology.

Okada, K., Venezia, J. H., Matchin, W., Saberi, K., & Hickok, G. (2013). An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex. *PLoS One*, *8*(6), 1–9. doi:10.1371/journal.pone.0068959

Olson, I. R., Gatenby, J. C., & Gore, J. C. (2002). A comparison of bound and unbound audio–visual information processing in the human cerebral cortex. *Cognitive Brain Research*, *14*(1), 129–138. doi:10.1016/S0926-6410(02)00067-8

Oray, S., Lu, Z. L., & Dawson, M. E. (2002). Modification of sudden onset auditory ERP by involuntary attention to visual stimuli. *International Journal of Psychophysiology*, *43*, 213–224. doi:10.1016/S0167-8760(01)00174-X

Ortega-Llebaria, M., Faulkner, A., & Hazan, V. (2001). Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English. In *AVSP 2001 International Conference on Auditory-Visual Speech Processing* (pp. 149–154). Tokyo, Japan.

Ostroff, W. L. (2000). *Non-Linguistic Influences on Infants' Nonnative Phoneme Perception: Exaggerated Prosody and Visual Speech Information Improve Discrimination*. Virginia Polytechnic Institute and State University.

Oyama, S. (1978). The sensitive period and comprehension of speech. *Working Papers on Bilingualism*, *16*, 1–17.

Paris, T., Kim, J., & Davis, C. (2013). Visual speech form influences the speed of auditory speech processing. *Brain and Language*, *126*(3), 350–6. doi:10.1016/j.bandl.2013.06.008

Pascalis, O., Haan, M. D., & Nelson, C. A. (2002). Is face processing species-specific during the first year of life? *Science*, *296*, 1321–1323.

Pascalis, O., Loevenbruck, H., Quinn, P. C., Kandel, S., Tanaka, J. W., & Lee, K. (2014). On the links among face processing, language processing, and narrowing during development. *Child Development Perspectives*, *8*, 65–70. doi:10.1111/cdep.12064

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., … Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, *10*. doi:10.1371/journal.pbio.1001251

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, *22*, 237–247. doi:10.1016/S0163-6383(99)00003-X

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information Michelle L . Patterson and Janet F . Werker. *Developmental Science*, *2*(6), 191–196.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, *36*, 767–776. doi:10.1016/S0896-6273(02)01060-7

Pegg, J. E., & Werker, J. F. (1997). Adult and infant perception of two English phones. *The Journal of the Acoustical Society of America*, *102*(6), 3742–53.

Penfield, W., & Roberts, L. (1959). *Penfield. Speech and Brain Mechanisms* (Princeton.). NJ : Princeton University Press.

Perani, D., Abutalebi, J., Paulesu, E., Brambati, S., Scifo, P., Cappa, S. F., & Fazio, F. (2003). The role of age of acquisition and language usage in early, high-proficient bilinguals: An fMRI study during verbal fluency. *Human Brain Mapping*, *19*, 170–182. doi:10.1002/hbm.10110

Pilling, M. (2009). Auditory Event-Related Potentials. *Journal of Speech, Language, and Hearing Research*, *52*, 1073–1082.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, *13*, 253–260. doi:10.3758/BF03214136

Pisoni, D. B. (1990). Effects of Talker Variability on Speech Perception: Implications for Current Research and Theory. In *ICSLP-1990* (pp. 1399–1408). Kobe, Japan.

Polivanov, E. (1931). La perception des sons d'une langue étrangère. *Travaux Du Cercle Linguistique de Prague*, *3*, 111–114.

Polka, L. (1991). Cross-language speech perception in adults : phonemic, phonetic, and acoustic contributions. *The Journal of the Acoustical Society of America*, *89*(6), 2961–2977.

Polka, L. (1992). Characterizing the influence of native language. *Perception & Psychophysics*, *52*(1), 37–52.

Polka, L. (1995). Linguistic influences in adult perception of non-native vowel contrasts. *The Journal of the Acoustical Society of America*, *97*(2), 1286. doi:10.1121/1.412170

Polka, L., & Bohn, O. S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. *The Journal of the Acoustical Society of America*, *100*(1), 577–92.

Polka, L., Colantonio, C., & Sundara, M. (2001). A cross-language comparison of /d/–/ð/ perception: Evidence for a new developmental pattern. *The Journal of the Acoustical Society of America*, *109*(5), 2190–2201. doi:10.1121/1.1362689

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology*, *20*(2), 421–35.

Pons, F., & Lewkowicz, D. J. (2014). Infant perception of audio-visual speech synchrony in familiar and unfamiliar fluent speech. *Acta Psychologica*, *149*, 142–7. doi:10.1016/j.actpsy.2013.12.013

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10598–602. doi:10.1073/pnas.0904134106

Preminger, J. E., Lin, H. B., Payen, M., & Levitt, H. (1998). Selective visual masking in speechreading. *Journal of Speech, Language, and Hearing Research*, *41*, 564–575.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(20), 7865–70. doi:10.1073/pnas.0509989103

Pulvermüller, F., & Schumann, J. (1994). Neurobiological mechanisms of language acquisition. *Language Learning*, *44*, 681–734.

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron*, *28*, 617–625. doi:10.1016/S0896-6273(00)00138-0

Ransdell, S. E., & Fischler, I. (1987). Memory in a Monolingual Mode : When Are Bilinguals a Disadvantage ? *Journal of Memory and Language*, *26*, 392–405.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-reading* (pp. 97–113). London: Lawrence Erlbaum Associates.

Remez, R. E. (1996). Critique: auditory form and gestural topology in the perception of speech. *The Journal of the Acoustical Society of America*, *99*, 1695–1698. doi:10.1121/1.414693

Remez, R. E. (2005). Three puzzles of multimodal speech perception. In E. Vatikiotis-Bateson, G. Bailly, & P. Perrier (Eds.), *Audiovisual Speech* (pp. 12–19). Cambridge, MA: MIT Press.

Ressel, V., Pallier, C., Ventura-Campos, N., Díaz, B., Roessler, A., Ávila, C., & Sebastián-Gallés, N. (2012). An effect of bilingualism on the auditory cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(47), 16597–601. doi:10.1523/JNEUROSCI.1996-12.2012

Ridouane, R. (2003). *Consonant Clusters in Berber: Phonetics and Phonology*. Université de la Sorbonne Nouvelle - Paris III, France.

Rivera-Gaxiola, M., Klarman, L., Garcia-Sierra, A., & Kuhl, P. K. (2005). Neural patterns to speech and vocabulary growth in American infants. *Neuroreport*, *16*, 495–498. doi:10.1097/00001756-200504040-00015

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–92. doi:10.1146/annurev.neuro.27.070203.144230

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*(2), 131–141. doi:10.1016/0926-6410(95)00038-0

Robert-Ribes, J., Schwartz, J.-L., Lallouache, T., & Escudier, J.-P. (1998). Complementarity and Synergy in Bimodal Speech : Auditory, visual, and audio-visual Identification of French Oral Vowels in Noise. *Journal of Acoustical Society of America*, *103*(6), 3677–3681. doi:10.1121/1.423069

Romaine, S. (1995). *Bilingualism*. (2nd ed.). Malden, MA: Blackwell.

Rönnberg, J. (1995). What makes a skilled speechreader ? In G. Plant & K.-E. Spens (Eds.), *Profound deafness and speech communication* (pp. 393–416). London: Whurr.

Rosenblum, L. D. (2005). Primacy of Multimodal Speech Perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 51–78). Oxford, UK: Blackwell Publishing.

Rosenblum, L. D. (2008). Primacy of Multimodal Speech Perception. In *The Handbook of Speech Perception* (pp. 51–78). doi:10.1002/9780470757024.ch3

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects: Research report. *Psychological Science*, *18*, 392–396. doi:10.1111/j.1467-9280.2007.01911.x

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*, 347–357. doi:10.3758/BF03211902

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007b). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–53. doi:10.1093/cercor/bhl024

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928. doi:10.1126/science.274.5294.1926

Saffran, J. R., Werker, J. F., & Werner, L. A. (2006). The Infant's Auditory World: Hearing, Speech, and the Beginnings of Language. In W. Damon & R. M. Lerner (Eds.), *Handbook of Child Psychology* (pp. 58–108). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception & Psychophysics*, *54*, 406–416. doi:10.3758/BF03205276

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V, Lu, S. T., & Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*(1), 141–5.

Sams, M., Möttönen, R., & Sihvonen, T. (2005). Seeing and hearing others and oneself talk. *Cognitive Brain Research*, *23*, 429–435. doi:10.1016/j.cogbrainres.2004.11.006

Sanders, L. D., Newport, E. L., & Neville, H. J. (2002). Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech. *Nature neuroscience*, 5, 700–703. doi:10.1038/nn873

Savariaux, C., Perrier, P., Orliaguet, J.-P., & Schwartz, J. (1999). Compensation for the perturbation of French [u] using a lip tube: II. Perceptual analysis. *Journal of Acoustic Society of America*, *106*(1), 381–393.

Scherg, M., & von Cramon, D. (1986). Psychoacoustic and electrophysiologic correlates of central hearing disorders in man. *European archives of psychiatry and neurological sciences,* 236, 56–60). doi:10.1007/BF00641060

Scholes, R. J. (1967). Phoneme Categorization of Synthetic Vocalic Stimuli By Speakers of Japanese, Spanish, Persian, and American English. *Language and Speech*, *10*(1), 46–68. doi:10.1177/002383096701000104

Scholes, R. J. (1968). Phonemic Interference as a Perceptual Phenomenon. *Language and Speech*, *11*(2). doi:10.1177/002383096801100202

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level, "unisensory" processing. *Current Opinion in Neurobiology*, *15*, 454–458. doi:10.1016/j.conb.2005.06.008

Schwartz, J. (2011). Analyse audiovisuelle des scènes de parole. In *Rencontres Jeunes Chercheurs en Parole, RJCP2011.* Grenoble, France.

Schwartz, J., Abry, C., Boë, L.-J., & Cathiard, M.-A. (2002). Phonology in a Theory of Perception-for-Action-Control. In J. Durand & B. Laks (Eds.), *Phonetics, Phonology, and Cognition* (pp. 254–280). Oxford, UK: Blackwell Publishing.

Schwartz, J., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *5*(25), 336–354.

Schwartz, J., Robert-Ribes, J., & Escudier, J.-P. (1998). Ten years after Summerfield. A Taxonomy of Models of Audiovisual Fusion in Speech Perception. In R. Campbell, B. Dodd, & D. K. Burnham (Eds.), *Hearing by Eye* (pp. 85–108). Hove, UK: Lawrence Erlbaum Associates Ltd.

Schwartz, J., Sato, M., & Fadiga, L. (2008). The common language of speech perception and action : a neurocognitive perspective. *Revue Française de Linguistique Appliquée*, *13*(2), 9–22.

Schwartz, J., & Savariaux, C. (2013). Data and simulations about audiovisual asynchrony and predictability in speech perception. In *12th International Conference on Auditory-Visual Speech Processing (AVSP 2013)* (pp. 147–152). Annecy, France.

Schwartz, J.-L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *The Journal of the Acoustical Society of America*, *127*, 1584–1594. doi:10.1121/1.3293001

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), 69–78. doi:10.1016/j.cognition.2004.01.006

Schwartz, J.-L., & Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology*, *10*(7), e1003743. doi:10.1371/journal.pcbi.1003743

Sebastián-Gallés, N., Albareda-Castellot, B., Weikum, W. M., & Werker, J. F. (2012). A bilingual advantage in visual language discrimination in infancy. *Psychological Science*, *23*(9), 994–9. doi:10.1177/0956797612436817

Sebastián-Gallés, N., Soriano-Mas, C., Baus, C., Díaz, B., Ressel, V., Pallier, C., … Pujol, J. (2012). Neuroanatomical markers of individual differences in native and non-native vowel perception. *Journal of Neurolinguistics*, *25*(3), 150–162. doi:10.1016/j.jneuroling.2011.11.001

Sekiyama, K., & Burnham, D. (2004). Issues in the Development of Auditory-Visual Speech Perception : Adults , Infants , and Children. In S. H. Kim & D. H. Youn (Eds.), *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing* (pp. 1137–1140). Jeju Island, South Korea.

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science*, *11*(2), 306–20. doi:10.1111/j.1467-7687.2008.00677.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research*, *47*(3), 277–287. doi:10.1016/S0168-0102(03)00214-1

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*, 1797–805.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, *21*, 427–444.

Sennema, A., Hazan, V., & Faulkner, A. (2003). The role of visual cues in L2 consonant perception. In *The International Congress of Phonetic Scences* (pp. 135–138). Barcelona, Spain.

Service, E. (1992). Phonology, working memory, and foreign-language learning. *The Quarterly Journal of Experimental Psychology*, *45*, 21–50. doi:10.1080/14640749208401314

Shahin, A. J., Kerlin, J. R., Bhat, J., & Miller, L. M. (2012). Neural restoration of degraded audiovisual speech. *NeuroImage*, *60*(1), 530–538. doi:10.1016/j.neuroimage.2011.11.097.Neural

Singleton, D. M., & Lengyel Zsolt. (1995). *The Age Factor in Second Language Acquisition: A Critical Look at the Critical Period Hypothesis*. (D. Singleton & Lengyel Zsolt, Eds.) (p. 157). Clevedon: Multilinguals Matters.

Skoruppa, K., Pons, F., Christophe, A., Bosch, L., Dupoux, E., Sebastián-Gallés, N., … Peperkamp, S. (2009). Language-specific stress perception by 9-month-old French and Spanish infants. *Developmental Science*, *12*(6), 914–919. doi:10.1111/j.1467-7687.2009.00835.

Slater, A. M., Quinn, P. C., Brown, E., & Hayes, R. (1999). Intermodal perception at birth: Intersensory redundancy guides newborn infants' learning of arbitrary auditory-visual pairings. *Developmental Science*, *2*, 333–338. doi:10.1111/1467-7687.00079

Sliwa, J., Duhamel, J.-R., Pascalis, O., & Wirth, S. (2011). Spontaneous voice-face identity matching by rhesus monkeys for familiar conspecifics and humans. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(4), 1735–40. doi:10.1073/pnas.1008169108

Smeele, P. M. T. (1994). *Perceiving Speech : Integration Auditory and Visual Speech*. Delft University of Technology, Pays-Bas.

Smeele, P. M. T., & Sitting, A. C. (1991). The contribution of Vision to Speech Perception. In *Proceedings of the Second Europen Conference on Speech Communication and Technology Eurospeech 91* (pp. 1495–1497). Geneva, Italia.

Smits, R. (2000). Temporal distribution of information for human consonant recognition in VCV utterances. *Journal of Phonetics*, *28*, 111–135. doi:10.006/jpho.2000.0107

Smits, R., Warner, N., McQueen, J. M., & Cutler, A. (2003). Unfolding of phonetic information over time: a database of Dutch diphone perception. *The Journal of the Acoustical Society of America*, *113*(1), 563–74.

Snow, C. E., & Hoefnagel-höhle, M. (1978). The Critical Period for Language Acquisition : Evidence from Second Language Learning. *Child Development*, *49*(4), 1114–1128.

Stein, B. E., & Meredith, M. A. (1993). *The Merging Senses*. *Cognitive Neuroscience Series* (p. 221). Cambridge, MA: MIT Press.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–73. doi:10.1162/jocn.2007.19.12.1964

Stevens, K. N. (1975). The potential role of property detectors in the perception of consonants. In G. Fant & M. A. Tatham (Eds.), *Auditory Analysis and Perception of Speech.* NewYork: Academy Press.

Strange, W. (1986). Speech input and the development of speech perception. In J. F. Kavanagh (Ed.), *Otitis media and child development* (pp. 12–26). Parkton, MD : Yorkton Press. doi:10.3840/000143

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *The Journal of the Acoustical Society of America*, *74*, 695–705. doi:10.1121/1.397863

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*, 212–215. doi:10.1121/1.1907309

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-reading* (pp. 3–51). London: Lawrence Erlbaum Associates.

Summerfield, Q. (1991). Visual perception of phonetic gestures. In I. G. Mattingly & M. Suddert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception* (pp. 117–137). Hillsdale: Erlbaum Associates.

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology*, *36*, 51–74. doi:10.1080/14640748408401503

Sundara, M., Polka, L., & Genesee, F. (2006). Language-experience facilitates discrimination of /d-th/ in monolingual and bilingual acquisition of English. *Cognition*, *100*, 369–388. doi:10.1016/j.cognition.2005.04.007

Tabouret-Keller, A. (2004). Bilinguism in Europe. In T. K. Bhatia & W. C. Ritchie (Eds.), *The Handbook of Bilingualism* (pp. 662–688). Oxford, UK: Blackwell Publishing Ltd. doi:10.1002/9780470756997

Takagi, N., & Mann, V. A. (1995). The limits of extended naturalistic exposure on the perceptual mastery of English /r/ and /l/ by adult Japanese learners of Englis. *Applied Psycholinguistics*, *16*(4), 379–405.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*(3), 850–5. doi:10.1016/j.cognition.2008.05.009

Terbeek, D. (1977). A Cross-Language Multidimensional Scaling Study of Vowel Perception. *University of California Working Papers in Phonetics*, *37*, 1–271.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. (L. B. Smith & E. Thelen, Eds.)*Journal of Cognitive Neuroscience* (Vol. 512, p. 376). Cambridge, MA: MIT Press. doi:10.1162/jocn.1995.7.4.512

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J.-R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(11), 4567–70. doi:10.1073/pnas.0811155106

Thomas, S. M., & Jordan, T. R. (2004). Contributions of Oral and Extraoral Facial Movement to Visual and Audiovisual Speech Perception. *Journal of Experimental Psychology*, *30*(5), 873–888. doi:10.1037/0096-1523.30.5.873

Thompson, M., & Hazan, V. (2010). The impact of visual cues and lexical knowledge on the perception of a non-native consonant contrast for Colombian adults. In K. Dziubalska-Kołaczyk, M. Wrembel, & M. Kum (Eds.), *Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech, New Sounds 2010* (pp. 493–498). Poznan, Poland.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522. doi:10.1038/381520a0

Tomé, M. (1997). La perception de los sonidos del francés por los estudiantes españoles. *Estudios Humanisticos - Filologia*, *19*, 263–269.

Treille, A., Cordeboeuf, C., Vilain, C., & Sato, M. (2014). Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions. *Neuropsychologia*, *57*, 71–7. doi:10.1016/j.neuropsychologia.2014.02.004

Troille, E., Cathiard, M.-A., & Abry, C. (2007). Consequences on bimodal perception of the timing of the consonant and vowel audiovisual flows. *Auditory-Visual Speech Processing*, 36–36.

Troille, E., Cathiard, M.-A., & Abry, C. (2010). Speech face perception is locked to anticipation in speech production. *Speech Communication*, *52*(6), 513–524. doi:10.1016/j.specom.2009.12.005

Troubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. *Travaux du cercle linguistique de Prague 7 : tard. Principes de phonologie* (1949th ed.), (pp. 47–47). Paris: Klincksieck. doi:10.1080/00233608808604171

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007a). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing*, *28*, 656–668. doi:10.1097/AUD.0b013e31812f7185

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007b). The effects of age and gender on lipreading abilities. *Journal of the American Academy of Audiology*, *18*(10), 883–92.

Vallabha, G. K., & McClelland, J. L. (2007). Success and failure of new speech category learning in adulthood: consequences of learned Hebbian attractors in topographic maps. *Cognitive, Affective & Behavioral Neuroscience*, *7*, 53–73. doi:10.3758/CABN.7.1.53

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. doi:10.1016/j.neuropsychologia.2006.01.001

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, *60*, 926–940.

Ventura-Campos, N., Sanjuán, A., González, J., Palomar-García, M.-Á., Rodríguez-Pujadas, A., Sebastián-Gallés, N., … Ávila, C. (2013). Spontaneous brain activity predicts learning ability of foreign sounds. *The Journal of Neuroscience*, *33*(22), 9295–305. doi:10.1523/JNEUROSCI.4655-12.2013

Von Holzen, K., & Mani, N. (2012). Language nonselective lexical access in bilingual toddlers. *Journal of Experimental Child Psychology*, *113*, 569–586. doi:10.1016/j.jecp.2012.08.001

Walley, A. C., & Flege, J. E. (1999). Effects of lexical status on children's and adults' perception of native and non-native vowels. *Journal of Phonetics*, *27*, 307–332. doi:10.1121/1.406508

Walsh, M. A., & Diehl, R. L. (1991). Formant transition duration and amplitude rise time as cues to the stop/glide distinction. *The Quarterly Journal of Experimental Psychology*, *43*, 603–620. doi:10.1080/14640749108400989

Walton, G. E., & Bower, T. G. R. (1993). Amodal representation of speech in infants. *Infant Behavior and Development*, *16*(2), 233–243. doi:10.1016/0163-6383(93)80019-5

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), 1716–26. doi:10.1121/1.2956483

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics*, *37*(3), 344–356. doi:10.1016/j.wocn.2009.04.002

Wassenhove, V. Van, Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–1186.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science*, *316*(5828), 1159. doi:10.1126/science.1137686

Werker, J. F., Frost, P. E., & McGurk, H. (1992). La langue et les Lèvres: cross-language influences on bimodal speech perception. *Canadian Journal of Psychology*, *46*(4), 551–68.

Werker, J. F., Gilbert, J. H., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development*, *52*, 349–355. doi:10.2307/1129249

Werker, J. F., & Tees, R. C. (1984). Cross-Language Speech Perception: Evidence for Perceptual Reorganization During the First Year of Life. *Infant Behavior and Development*, (7), 49–63.

Whitley, M. S. (2002). *Spanish/English Contrasts: A Course in Spanish Linguistics* (p. 402). Washington, DC: Georgetown University Press.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*(7), 701–2. doi:10.1038/nn1263

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex*, *13*, 1034–1043. doi:10.1093/cercor/13.10.1034

Yakel, D. A. (2000). *Effects of Time-Varying Information on Vowel Identification Accuracy in Visual Speech Perception*. *Dissertation Abstracts International, B: Sciences and Engineering*. University of California, Riverside.

Yi, H.-G., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America*, *134*(5), EL387–93. doi:10.1121/1.4822320

Zhang, Y., Kuhl, P. K., Imada, T., Iverson, P., Pruitt, J., Stevens, E. B., … Nemoto, I. (2009). Neural signatures of phonetic learning in adulthood: A magnetoencephalography study. *NeuroImage*, *46*, 226–240. doi:10.1016/j.neuroimage.2009.01.028

Zhang, Y., Kuhl, P. K., Imada, T., Kotani, M., & Tohkura, Y. (2005). Effects of language experience: neural commitment to language-specific auditory patterns. *NeuroImage*, *26*(3), 703–20. doi:10.1016/j.neuroimage.2005.02.040

Zion Golumbic, E., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual Input Enhances Selective Speech Envelope Tracking in Auditory Cortex at a "Cocktail Party." *Journal of Neurosciences*, *33*(4), 1417–1426. doi:10.1523/JNEUROSCI.3675-12.2013.Visual

LISTE DES FIGURES

LISTE DES TABLEAUX

ANNEXE 1

Code sujet :

Partie A

- Date: _____

- Date de naissance : _____ - Lieu de naissance : _____

- Nationalité : _____ - ☐ Masculin ☐ Féminin

A1. Quelles sont les langues que tu connais et quel est l'ordre dans lequel tu les **maîtrises** ? *(langue la mieux maîtrisée en premier)*

1	**2**	**3**	**4**	**5**

A2. Dans quel **ordre** as-tu acquis ces langues ? *(langue maternelle en premier)*

1	**2**	**3**	**4**	**5**

A3. Quel est le pourcentage moyen de temps au cours duquel tu es **exposé actuellement** à chacune de tes langues ? *(la somme doit donner 100%)*

Langue					
Pourcentage					

A4. Imagine que tu as la possibilité de **lire un texte** disponible dans toutes les langues qui te sont familières. Dans quelle mesure (en pourcentage) choisis-tu de le lire dans chacune de tes langues (si l'original a été écrit dans une langue qui t'est inconnue) ? *(la somme doit donner 100%)*

Langue					
Pourcentage					

A5. Imagine que tu engages une **conversation** avec une personne qui parle chacune de tes langues. Dans quelle mesure (en pourcentage) choisis-tu d'employer chacune de tes langues ? *(la somme doit donner 100%)*

Langue					
Pourcentage					

A6. As-tu déjà **vécu** dans un pays autre que la France, et si oui, quand et pour combien de temps ?

Pays : _____ Dates : _____ Durée : _____

Pays : _____ Dates : _____ Durée : _____

Date depuis laquelle tu habites en France : _____

Partie B

(remplir un questionnaire B pour chaque langue)

Langue : _____

Toutes les questions ci-dessous se réfèrent à ta connaissance de cette langue.

B1. A quel âge as-tu commencé à apprendre cette langue ? _____

à parler couramment cette langue ? _____

à lire dans cette langue ? _____

à lire couramment dans cette langue ? _____

à écrire couramment dans cette langue ? _____

B2. Combien de temps as-tu passé dans les environnements linguistiques suivants ?

	Années	**Mois**
Pays où la langue est parlée :		
Famille où la langue est parlée :		
Etablissement scolaire où cette langue est la principale langue d'enseignement :		

B3. Sur une échelle de 0 à 10, quelle est ta **compétence** dans cette langue dans les domaines suivants ? *(0=aucune, 1=très faible, 5=satisfaisante, 9=très bonne, 10=excellente)*

	0	1	2	3	4	5	6	7	8	9	10
Expression orale :	O	O	O	O	O	O	O	O	O	O	O
Compréhension orale :	O	O	O	O	O	O	O	O	O	O	O
Lecture :	O	O	O	O	O	O	O	O	O	O	O
Écriture :	O	O	O	O	O	O	O	O	O	O	O

B4. Selon toi, sur une échelle de 0 à 10, dans quelle mesure les facteurs suivants ont-ils contribués à ton apprentissage de cette langue ? *(0=pas du tout, 1= minimalement, 5=moyennement, 9= beaucoup, 10=très fortement)*

Code sujet :

	0	1	2	3	4	5	6	7	8	9	10
Interaction avec des amis :	O	O	O	O	O	O	O	O	O	O	O
Interaction avec ta famille :	O	O	O	O	O	O	O	O	O	O	O
Apprentissage à l'école :	O	O	O	O	O	O	O	O	O	O	O
Personne/s qui t'a/ont gardé lorsque tu étais petit/e :	O	O	O	O	O	O	O	O	O	O	O
Auto-apprentissage :	O	O	O	O	O	O	O	O	O	O	O
Lecture :	O	O	O	O	O	O	O	O	O	O	O
Regarder la télévision ou des films :	O	O	O	O	O	O	O	O	O	O	O
Ecouter la radio ou de la musique :	O	O	O	O	O	O	O	O	O	O	O

B5. Dans quelle mesure es-tu **actuellement exposé** à cette langue dans les contextes suivants ? *(0= jamais dans ce contexte, 2=presque jamais, 5=la moitié du temps, 8=la plupart du temps, 10=tout le temps)*

	0	1	2	3	4	5	6	7	8	9	10
Interaction avec des amis :	O	O	O	O	O	O	O	O	O	O	O
Interaction avec ta famille :	O	O	O	O	O	O	O	O	O	O	O
Auto-apprentissage :	O	O	O	O	O	O	O	O	O	O	O
Lecture :	O	O	O	O	O	O	O	O	O	O	O
Écriture :	O	O	O	O	O	O	O	O	O	O	O
Regarder la télévision	O	O	O	O	O	O	O	O	O	O	O

ou des films :											
Ecouter la radio ou de la musique :	O	O	O	O	O	O	O	O	O	O	O

B6. Selon toi, à quel point as-tu un **accent étranger** dans cette langue ? *(0=aucun accent, 1=presque aucun, 5=modéré, 9=plutôt fort, 10=extrêmement fort)*

	0	1	2	3	4	5	6	7	8	9	10
Accent étranger :	O	O	O	O	O	O	O	O	O	O	O

B7. Dans quelle mesure d'autres personnes reconnaissent à ton accent que tu n'es pas un locuteur natif ? (*0= jamais, 1=presque jamais, 5=la moitié du temps, 9=presque toujours, 10=toujours)*

	0	1	2	3	4	5	6	7	8	9	10
	O	O	O	O	O	O	O	O	O	O	O

ANNEXE 2

Burfin, S., Pascalis, O., Ruiz Tada, E., Costa, A., Savariaux, C. & Kandel, S. (2014)

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 21 October 2014
doi: 10.3389/fpsyg.2014.01179

Bilingualism affects audiovisual phoneme identification

Sabine Burfin[1], Olivier Pascalis[1], Elisa Ruiz Tada[2], Albert Costa[2,3], Christophe Savariaux[4] and Sonia Kandel[1,4,5]*

[1] *LPNC (CNRS UMR 5105) – Université Grenoble Alpes, Grenoble, France*
[2] *Universitat Pompeu Fabra, Barcelona, Spain*
[3] *Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain*
[4] *GIPSA-lab (CNRS UMR 5216) – Université Grenoble Alpes, Grenoble, France*
[5] *Institut Universitaire de France*

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 21 October 2014
doi: 10.3389/fpsyg.2014.01179

Bilingualism affects audiovisual phoneme identification

Sabine Burfin[1], Olivier Pascalis[1], Elisa Ruiz Tada[2], Albert Costa[2,3], Christophe Savariaux[4] and Sonia Kandel[1,4,5]*

[1] LPNC (CNRS UMR 5105) – Université Grenoble Alpes, Grenoble, France
[2] Universitat Pompeu Fabra, Barcelona, Spain
[3] Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain
[4] GIPSA-lab (CNRS UMR 5216) – Université Grenoble Alpes, Grenoble, France
[5] Institut Universitaire de France

Edited by:
Noel Nguyen, Université d'Aix-Marseille, France

Reviewed by:
Anahita Basirat, Lille 2 University, France
Kauyumari Sanchez, New Zealand Institute of Language, Brain and Behaviour, New Zealand

***Correspondence:**
Sonia Kandel, Laboratoire de Psychologie et NeuroCognition (CNRS UMR 5105), Université Grenoble 2, BSHM, BP 47, 38040 Grenoble 09, France
e-mail: sonia.kandel@upmf-grenoble.fr

We all go through a process of perceptual narrowing for phoneme identification. As we become experts in the languages we hear in our environment we lose the ability to identify phonemes that do not exist in our native phonological inventory. This research examined how linguistic experience—i.e., the exposure to a double phonological code during childhood—affects the visual processes involved in non-native phoneme identification in audiovisual speech perception. We conducted a phoneme identification experiment with bilingual and monolingual adult participants. It was an ABX task involving a Bengali dental-retroflex contrast that does not exist in any of the participants' languages. The phonemes were presented in audiovisual (AV) and audio-only (A) conditions. The results revealed that in the audio-only condition monolinguals and bilinguals had difficulties in discriminating the retroflex non-native phoneme. They were phonologically "deaf" and assimilated it to the dental phoneme that exists in their native languages. In the audiovisual presentation instead, both groups could overcome the phonological deafness for the retroflex non-native phoneme and identify both Bengali phonemes. However, monolinguals were more accurate and responded quicker than bilinguals. This suggests that bilinguals do not use the same processes as monolinguals to decode visual speech.

Keywords: phonological deafness, bilinguals, monolinguals, audiovisual speech

INTRODUCTION

Having a conversation in a non-native language is a difficult task when our proficiency with that language is limited. To have a fluent conversation, we have to learn new vocabulary, syntax rules, and deal with new speech sounds. The present study examined whether the visual information provided by the articulatory gestures of the speaker enhances the identification of phonemes of other languages, especially those that do not exist in our phonological inventory. For example, the English phoneme /θ/ does not exist in French. Native French speakers have difficulties identifying it and often confuse the /θ/-/f/ or /θ/-/s/ contrasts like in the words /θɪn/ (*thin*), /fɪn/ (*fin*), and /sɪn/ (*sin*). So when they hear *thin*, they assimilate /θ/ to the closest phoneme they know, in this case to /f/ or /s/ (Best, 1995). This example illustrates the phenomenon of "phonological deafness" (Polivanov, 1931). It refers to the difficulty or inability to identify phonemes that do not exist in the native phoneme inventory. When the non-native phoneme shares phonetic features with phonemes that exist in our phonological repertoire, we tend to confuse the non-native phoneme with the native one. This research examined how linguistic experience—i.e., the exposure to a double phonological code during childhood—affects the visual processes involved in non-native phoneme identification.

Phonological deafness results from perceptual narrowing processes. At birth, infants are capable of discriminating all phonological contrasts (Werker and Tees, 1984; Kuhl et al., 2006). This ability decreases progressively during the first year of life. As we become experts in the languages we hear in our environment we lose the ability to identify phonemes that do not exist in our native phonological inventory. Werker and Tees (1984) showed that 6–10 months old English-speaking babies could discriminate Salish and Hindi consonant phonemes that do not exist in English. At 10 months, their ability to discriminate these phonemes decreased significantly and almost disappeared at 12 months. The 12-months old children were phonologically deaf to these phonemes whereas Salish and Hindi-speaking infants of the same age could distinguish the phonemes perfectly well. These experiments tested the infants on auditory perception. A more recent study provided evidence for perceptual narrowing also in audiovisual speech. In Spanish, the English contrast /b/-/v/ (like in the words *ban* and *van*) does not exist. Phoneme /v/ does not exist in Spanish and is often assimilated to /b/ that does exist. Pons et al. (2009) presented this contrast audiovisually to English and Spanish-speaking 6 and 11 months infants. The results showed that both groups were audiovisually sensitive to the /b/-/v/ contrast at 6 months. At 11 months however, the Spanish-speaking babies lost this sensitivity but not the English-speaking ones. Although this decrease in phoneme discrimination abilities is a well-known phenomenon (see Best, 1995 for a review) and constitutes a real difficulty for second language learners, it is clear that

we are all able to learn a non-native language after 12 months. A great majority of the world's population can communicate in several languages without having grown up in a multi-lingual environment (Altarriba and Heredia, 2008).

Most of us have experienced that to have a conversation in a non-native language is very difficult when we are not proficient with it. It becomes even more difficult if we cannot see the speaker's face, like when we are on the phone. This is likely due to the fact that on the phone we cannot see the articulatory movements of the speaker, which may provide visual cues on phoneme identity. A few studies presented data indicating that when having to deal with a non-native language, these visual cues may enhance performance (Davis and Kim, 2001; Hazan et al., 2006). Burfin et al. (2011) conducted an experiment that directly concerns phonological deafness. French native speakers had to identify the Spanish inter-dental fricative phoneme /θ/; that does not exist in French. The participants systematically identified /θ/ as /f/ when the phonemes were presented auditorily. In other words, they were phonologically deaf to the /θ/-/f/ contrast. In an audiovisual presentation, where the participants could see the speaker producing the phonemes, /θ/ was no longer confused with /f/. It was identified correctly up to 80–90%. This suggests that the participants used the visual cues provided by the speaker to overcome phonological deafness. This is in line with previous research presented by Navarra and Soto-Faraco (2007). They showed that Spanish-Catalan bilinguals who were Spanish dominant failed to distinguish the Catalan /e/-/ɛ/ contrast (that does not exist in Spanish) in an audio-only presentation. In contrast, they could discriminate the phonemes in an audiovisual presentation. Taken together, these studies, carried out in several languages, reveal that the visual information on the speaker's articulatory movements can be very useful to overcome—at least partially—phonological deafness.

The visual information on the speaker's speech movements also seems to play a key role to discriminate languages, but the visual sensitivity depends on early linguistic exposure. Weikum et al. (2007) conducted an experiment in which 6 and 8-months old infants viewed silent videos of a bilingual French-English speaker telling a story either in French or English. One group of infants lived in an English monolingual environment whereas the other grew up in a French-English bilingual environment. The results indicated that all the 6 month-old infants could distinguish the French and English stories. At 8 months-old the English monolingual group could not distinguish the English and French stories. The 8 months-old bilinguals instead could distinguish them. Sebastian-Gallés et al. (2012) provided further data indicating that the infants' linguistic experience at birth is determinant for developing the visual sensitivity to language discrimination, irrespective of the languages they are exposed to. The authors presented the same French and English silent video stories to 8 months monolingual Spanish or Catalan and bilingual Spanish-Catalan infants that had never heard English or French before. The results revealed that the bilingual group distinguished the English from French videos whereas the monolinguals did not. This suggests that monolinguals and bilinguals could use different processing mechanisms to decode visual speech. These results concerned language discrimination in a story. Do bilinguals process visual information for phonemes that do not exist in their phonological inventory as monolinguals do? The present study examined whether linguistic experience during early childhood affects the visual processes involved in non-native phoneme identification in audiovisual speech.

Children who grow up in a bilingual environment (or are exposed to a foreign language very early in life) seem to be particularly sensitive to native and non-native sounds. Byers-Heinlein et al. (2010) presented data indicating that newborns who were prenatally exposed to one language (i.e., English monolingual mother or Tagalog monolingual mother) preferred their mothers' language. If bilingual mothers spoke both languages during pregnancy, the newborns had no preference for either language. Furthermore, if the bilingual mothers spoke English and Chinese during pregnancy, the neonates had no preference for a language spoken during the pregnancy (English) or a new one (Tagalog). This suggests that "bilingual" neonates could process speech sounds differently (Burns et al., 2003; Kuhl et al., 2003) and lead to differences in the neural structure in the auditory cortex as adults (Ressel et al., 2012).

In sum, visual information on the face movements of the speaker seems to be extremely useful to decode speech. Monolinguals and bilinguals seem to process visual language differently, at least during the first year of life. Are these differences still present during adulthood? Do they use the same processing mechanisms to decode visual speech for phoneme identification? Do monolinguals and bilinguals use visual information to overcome phonological deafness? We conducted an experiment with monolingual and bilingual participants to answer these questions. The participants had to discriminate a Bengali plosive dental-retroflex contrast (/t/-/ʈ/) that does not exist in any of the participants' languages. The dental /t/ phoneme exists in all the participants' phonological inventories whereas the retroflex counterpart does not exist in any of them. The retroflex consonant we used is a coronal consonant where the tongue has a curled shape and is articulated between the alveolar ridge and the hard palate. The retroflex feature is articulated further back of the vocal tract than the dental. Moreover, during the articulation of the dental the tongue is apparent after the burst release. This means that the dental-retroflex contrast is both auditorily and visually salient. The recordings were presented in an audio-only condition to examine whether the two groups differed in their abilities to discriminate between native and non-native phonemes. We also presented the Bengali recordings with their corresponding videos in an audiovisual presentation to investigate whether the visual information on the speaker's face movements contributed to overcome the difficulties in phoneme identification for the non-native contrast. Since monolinguals and bilinguals process visual speech differently during early childhood, it is likely that their abilities for visual speech processing is also different as adults.

METHODS

PARTICIPANTS

Information on the participants' linguistic experience was collected with an adapted version of the "Language Experience and Proficiency Questionnaire" (Marian et al., 2007). There were 47

bilinguals. Although we did not directly study the issue of early vs. late bilingualism, we selected the bilingual participants on the basis of an early exposure to two languages and a high proficiency with them. There were 24 Catalan-Spanish bilinguals (4 men and 20 women; mean age = 20 years). They have all been exposed to both languages very early in childhood. Mean age of acquisition of Spanish and Catalan was 11 and 14 months, respectively. Ten learnt Spanish and Catalan at home. Seven have always been exposed to Spanish but lived in a Catalan environment and seven have been exposed to Catalan from birth but learnt Spanish in nursery school. They were students at the University Pompeu Fabra (Barcelona, Spain). There were 23 bilinguals of different languages (8 men and 15 women; mean age = 19.6 years). These bilinguals spoke French and another language from birth: English (4 participants), German (4 participants), Italian (4 participants), Spanish (3 participants), Malgash (2 participants), Portuguese (2 participants), Arab (2 participants), Polish (2 participants). They have all been exposed to both languages from birth. They spoke French because they lived in Grenoble from birth and the other language was their parents' native mother tongue. They were balanced bilinguals with equivalent proficiency in both languages (Marian et al., 2007). They were students at the University of Grenoble or at the Cité Scolaire Internationale which is an international school in Grenoble where only bilinguals can attend. The monolingual group consisted of 47 French native speakers (8 men and 39 women; mean age = 22 years). They all learnt English as a second language in middle and high school but their proficiency was very poor. They had no experience in a foreign country of more than 1 month. They were students at the University of Grenoble and received course credit for participation.

MATERIAL

We recorded 20 tokens of two Bengali syllables that differed in the plosive dental/retroflex phonological contrast (/ta/ and /ʈa/). This contrast does not exist in any of the languages spoken by the participants. The plosive dental phoneme /ta/ exists in all of the participants' languages, whereas the retroflex plosive /ʈa/ does not. The stimuli were recorded in a sound proof room by a native female Bengali speaker from Bangladesh. We presented the full face of the speaker with a blue background (see **Figure 1**). The recordings were done with a tri-CCD SONY DXC-990P camera and an AKG C1000S microphone. They were converted into AVI video files (PAL format, 25 img/s) and were segmented manually with the DpsRealitysoftware. Each /ta/ and /ʈa/ sequence began and ended with the speaker with the mouth closed. We selected 11 tokens of each sequence out of the 20 recordings. **Figure 1** describes the audio and visual characteristics of the stimuli for a dental and a retroflex token.

PROCEDURE

The experiment was conducted with an ABX paradigm. It was programmed with Eprime® software2.0 (Psychology Software Tools, Inc.). Ten tokens of /ta/ and 10 of /ʈa/ were used as X stimuli and one token of each syllable as references for A and B, for a total of 20 trails. We presented the A token, then the B token and finally the X stimulus. The participants were instructed to press one key of the keyboard if the X stimulus presented the same syllable as the A token or another key if the X stimulus presented the same syllable as the B token. The correct responses were equally distributed between both hands. The participants were instructed to respond as quickly and as accurate as possible. The X stimuli were presented randomly within a block. The identity (i.e., dental or retroflex) of the A and B tokens were fixed all along the experiment and controlled between participants. In the audio-only condition (A hereafter) we presented the audio track of the stimuli and the computer screen displayed an image with a still face of the speaker. In the audiovisual condition (AV hereafter) we presented the same audio track but the screen displayed the video

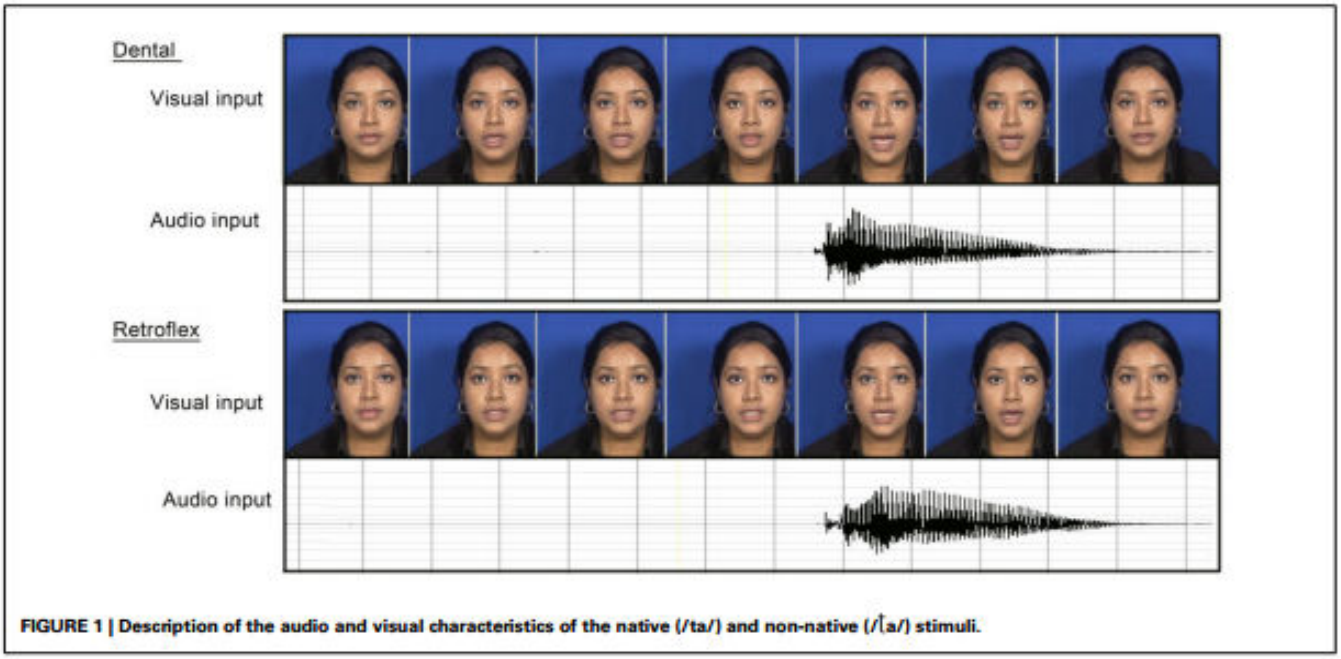

FIGURE 1 | Description of the audio and visual characteristics of the native (/ta/) and non-native (/ʈa/) stimuli.

with the moving face of the speaker. The experiment consisted of two blocks that were counterbalanced between participants; one for each presentation modality. The participant either heard (A-only) or saw and heard (AV condition) the Bengali sequences, or vice-versa.

The task was displayed by a Monitor LCD Dell (17 inches). The video stimuli were presented at 25 frames/s with a resolution of 720 × 576 pixels. The auditory component of the stimuli was provided at a 44100 Hz sampling rate by two SONY SRS-88 speakers located on both sides of the screen. In both conditions, the participants were instructed to respond on the basis of what they *perceived*, without referring to the auditory or visual modalities. We recorded correct responses (Accuracy) and the reaction time of the correct responses (RT). The participants were tested individually in a quiet room. They sat 40 cm away from the screen and the sound level was set to a comfortable level. Before the task, we made sure the participants understood the task by a short training session in A-only and AV presentation of Vietnamese consonant-vowel syllables. The experiment lasted approximately 30 min (questionnaires, instructions, and experimentation).

RESULTS

The results were analyzed using linear mixed effects models (Bates, 2005; Baayen et al., 2008), which simultaneously take participant and item variability into account. These analyses were performed using the software R (R Development Core Team, Bates and Maechler, 2009) with the package lme4 (Bates and Maechler, 2009). The statistical analyses were performed on Accuracy and Reaction time with Group (Monolingual, Bilingual), Modality (Audio-only, Audiovisual), and Phoneme (Native, Non-Native) as factors.

ACCURACY

The results for Accuracy are presented in **Table 1**.

The analysis revealed no significant main effects (**Table A1**). In contrast, the interactions between the factors were significant. The three way interaction was not significant, $t_{(3757)} = -1.14$, $p = 0.25$. The interaction between Modality and Group reached significance, $t_{(3757)} = 3.13$, $p < 0.001$. **Figure 2** presents the percentage of correct responses for monolinguals and bilinguals for the audio-only and audiovisual presentations.

Pairwise comparisons revealed that both groups had better scores in the AV than A presentations [monolinguals, $t_{(1879)} = 11.34$, $p < 0.001$; bilinguals, $t_{(1879)} = 6.07$, $p < 0.001$]. However, the "Audiovisual benefit" (AV score—A score) was higher for monolinguals (22%) than bilinguals (13%), $t_{(1878)} = 2.71$, $p < 0.01$. In the AV condition, monolinguals had higher scores than bilinguals, $t_{(1879)} = 3.50$, $p < 0.001$. In contrast, there were no group differences in the A condition, $t_{(1879)} = 0.22$, $p = 0.82$. To test for phonological deafness, we compared the accuracy scores for the non-native retroflex phoneme in the audio-only condition with chance level (50%) for each group. The scores for monolinguals (54.8%) did not reach significance [$T_{(1, 46)} = 1.85$, $p = 0.06$] and was slightly above chance (59.3%) for bilinguals, $T_{(1, 46)} = 3.04$, $p < 0.001$.

The interaction between Modality and Phoneme was significant, $t_{(3757)} = 3.08$, $p < 0.01$. **Figure 3** presents the percentage of correct responses for native and non-native phonemes for the audio-only and audiovisual presentations.

Pairwise comparisons revealed that for both phonemes the scores improved in the AV presentation with respect to the A presentation [native, $t_{(1879)} = 11.83$, $p < 0.001$; non-native, $t_{(1879)} = 6.20$, $p < 0.001$]. However, the "Audiovisual benefit" (AV score—A score) was greater for the native phonemes (22%) than non-native phonemes (13%), $t_{(1878)} = 3.23$, $p < 0.001$. In the audiovisual presentation, the scores for the native phonemes were higher than non-native phonemes, $t_{(1879)} = 4.90$, $p < 0.001$. In contrast, there were no differences between the phonemes in the audio-only condition, $t_{(1879)} = 0.84$, $p = 0.40$.

The interaction between Group and Phoneme was also significant, $t_{(3757)} = 2.42$, $p < 0.01$. **Figure 4** presents the percentage

Table 1 | Accuracy values (%) and standard errors (in brackets) for bilinguals and monolinguals in the A-only and AV presentation modalities for native and non-native phonemes.

	A-only		AV	
	Native	Non-native	Native	Non-native
Monolingual	62 (7)	55 (7)	88 (5)	75 (6)
Bilingual	56 (7)	59 (7)	76 (6)	66 (7)

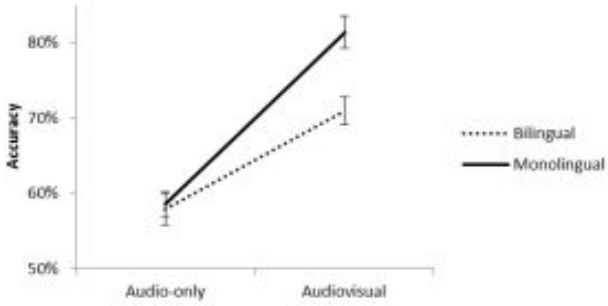

FIGURE 2 | Accuracy for monolinguals and bilinguals in the audio-only and audiovisual presentations of the Bengali phonemes.

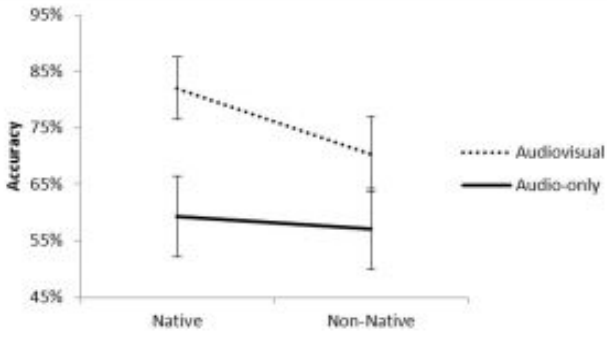

FIGURE 3 | Accuracy for native and non-native phonemes in the audio-only and audiovisual presentations.

of correct responses for native and non-native phonemes for monolinguals and bilinguals.

Pairwise comparisons revealed that the scores for the two groups were equivalent for the non-native phonemes, $t_{(1879)} = 0.64$, $p = 0.52$. In contrast, for the native phonemes the scores for monolinguals were higher than bilinguals, $t_{(1879)} = 2.40$, $p < 0.01$. For monolinguals, the scores for native phonemes were higher than the non-native ones, $t_{(1879)} = 4.34$, $p < 0.001$. For bilinguals, the scores were equivalent for the two kinds of phonemes, $t_{(1879)} = 1.61$, $p = 0.10$[1].

REACTION TIME

RTs faster than 300 ms and slower than 3000 ms were excluded from the analysis (4.41% of the data). The results for Reaction Time of correct responses are presented in **Table 2**.

The analysis (**Table A2**) revealed that monolinguals were globally faster than bilinguals, $t_{(3591)} = -2.45$, $p < 0.01$. The Modality effect almost reach significance, $t_{(3591)} = -1.82$, $p = 0.06$, indicating that correct responses were faster in the AV than A presentations. Responses for native phonemes were faster than non-native phonemes, $t_{(3591)} = -2.31$, $p < 0.05$. The three way interaction was not significant, $t_{(3757)} = 0.87$, $p = 0.38$. The interaction between Modality and Group was significant, $t_{(3591)} = -2.75$, $p < 0.001$. **Figure 5** presents the reaction time for monolinguals and bilinguals for the audio-only and audiovisual presentations.

Pairwise comparisons revealed that both groups respond faster in the audiovisual presentation with respect to the audio-only presentation [monolinguals, $t_{(1822)} = -10.22$, $p < 0.001$; bilinguals, $t_{(1822)} = -4.44$, $p < 0.001$]. The "Audiovisual benefit" (A RT—AV RT) was numerically greater for monolinguals (150 ms) than bilinguals (104 ms), but the difference did not reach significance, $t_{(1878)} = -1.00$, $p = 0.31$. In the audiovisual presentation, monolinguals were faster to respond than bilinguals, $t_{(1791)} = -3.20$, $p < 0.001$. In contrast, there were no group differences in the audio-only condition, $t_{(1791)} = -1.39$, $p = 0.16$.

DISCUSSION

The aim of this study was to examine whether monolinguals and bilinguals—who had a different linguistic experience during early childhood and after—take advantage of the visual information on the speaker's face movements when they have to identify phonemes that do not exist in their native language/s. Monolingual and bilingual participants had to identify Bengali phonemes that differ on a dental/retroflex consonant contrast that does not exist in any of the languages they speak. The phonemes were presented in an audiovisual condition where a video presented the visual information of the speaker producing the phonemes and their corresponding sound. In the audio-only condition the participants were presented the same stimuli but without the visual information. The results indicated that in the audio-only presentation monolinguals and bilinguals had similar difficulties in discriminating the Bengali retroflex phoneme. In the audiovisual condition instead, both groups took advantage of the visual information on the speaker's face movements to identify the non-native retroflex phoneme. They could overcome—at least partially- the difficulties experienced in the audio-only condition. The visual information not only enhanced identification but also accelerated phoneme processing. The results also point to visual processing differences between the two groups, since in the AV condition monolinguals were more accurate and faster than bilinguals. Furthermore, the "audiovisual benefit" was greater for monolinguals than bilinguals indicating that linguistic exposure

[1] We conducted an additional analysis to see whether the participants' performance increased in the second block: A then AV and AV then A. We did a LMM analysis with group (monolinguals, bilinguals) and presentation order (A then AV; AV then A) as factors. The results did not yield a significant block effect, $t_{(3758)} = 1.44$, $p = 0.14$.

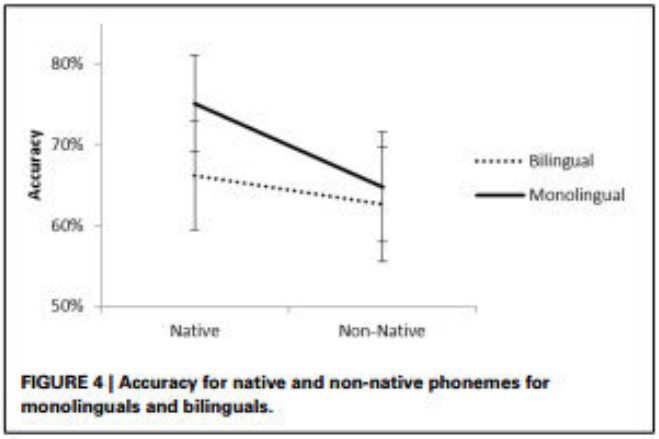

FIGURE 4 | Accuracy for native and non-native phonemes for monolinguals and bilinguals.

Table 2 | Reaction time values (ms) and standard errors (in brackets) for bilinguals and monolinguals in the A-only and AV presentation modalities for native and non-native phonemes.

	A-only		AV	
	Native	Non-native	Native	Non-native
Monolingual	1428 (70)	1498 (73)	1251 (62)	1303 (63)
Bilingual	1495 (70)	1573 (72)	1385 (73)	1501 (72)

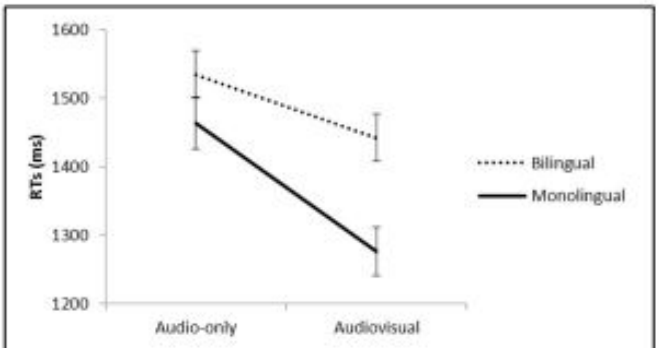

FIGURE 5 | Mean reaction times (ms) on correct responses for monolinguals and bilinguals in the audio-only and audiovisual presentations of the Bengali phonemes.

to more than one language may affect the visual processing of non-native phonemes.

Monolinguals and bilinguals had similar accuracy scores and reaction times in the audio-only condition. This suggests that when the participants heard the Bengali /ʈ/ retroflex phoneme—that does not exist in their phonological repertoire—they assimilated it to /t/ that exists in the languages they speak. This assimilation phenomenon leads to serious difficulties in phoneme identification and occurs because they cannot process the auditory relevant cue that distinguishes the two phonemes (Best et al., 2001). This is in agreement with previous research showing that early bilinguals can have monolingual-like performance during unfamiliar phoneme perception (Pallier et al., 1997; Sebastián-Gallés and Soto-Faraco, 1999). Also, Von Holzen and Mani (2012) showed that French-German preschooler bilinguals failed at discriminating Salish consonants. Moreover, Navarra and Soto-Faraco (2007)'s study indicated that in an audio-only condition Spanish-Catalan bilinguals who were Spanish dominant could not discriminate the Catalan phonemes /e/ and /ɛ/ (only /e/ exists in Spanish). This suggests that the particular sensitivity bilinguals have during the first year of life does not necessarily extend to non-native phonemes later in life. Even if several studies on infant perception showed more phoneme sensitivity in bilinguals, perceptual development keeps changing after the first year of life (Sundara et al., 2006). Our findings are in line with this idea, since in the audio-only condition the bilinguals did not exhibit a particular phoneme identification advantage with respect to monolinguals.

The main contribution of the present study concerns the visual component of non-native phoneme identification processes. All the participants could exploit the visual cues that distinguish retroflex /ʈ/ from dental /t/ in the audiovisual condition. Even if the retroflex feature does not exist in the participants' phonological inventory, the visual differences between the two consonants are salient enough to identify them. So the visual information on the speakers' facial movements played an important role in overcoming, at least partially, the phoneme identification difficulties the two groups experienced in the audio-only condition. The audio-visual benefit was higher for the native than the non-native phoneme. To our knowledge, the only study on audiovisual phoneme perception conducted with bilinguals is the one carried out by Navarra and Soto-Faraco (2007). As mentioned above, Spanish-Catalan bilinguals who were Spanish dominant could not discriminate the Catalan phonemes /e/ and /ɛ/ in an audio-only condition. In the AV condition, they were able to overcome the difficulties in phoneme identification, as in our experiment. What we do not know from Navarra and Soto-Faraco (2007) is whether these bilingual participants performed differently than monolinguals because the authors did not include a monolingual group in their study. Our research provides an answer to this question.

The results of the present study revealed that monolinguals and bilinguals do not take the same advantage of visual information on phoneme identity. In the AV condition, monolinguals were more accurate than bilinguals. In addition, the "Audiovisual benefit"—i.e., the accuracy score increase from A to AV—was 9% higher for monolinguals than bilinguals. We also observed that perceiving the speaker's speech gestures accelerated phoneme processing for both groups but again, the bilingual group seemed less sensitive to visual information. The acceleration was more pronounced in monolinguals than bilinguals. The difference in "Audiovisual benefit"—i.e., the decrease in reaction time—between the groups was of 46 ms but the difference failed to reach statistical significance. This is consistent with Sebastian-Gallés et al. (2012)'s study suggesting that monolinguals and bilinguals use different processing mechanisms to decode visual speech. However, the latter was conducted with 8 month-old infants and concerned language discrimination tasks and not phoneme identification, so we do not know which component of visual speech processing is responsible for these differences.

Studies on non-native phoneme identification/discrimination showed that early exposure to several languages delays perceptual narrowing and could lead to a better performance for non-native phonemes (Burns et al., 2003; Kuhl et al., 2003; Byers-Heinlein et al., 2010). Our findings do not confirm this "Bilingual advantage." Being exposed to several languages may delay perceptual narrowing during infancy, but it does not necessarily lead to a benefit for identifying non-native phonemes in adulthood. In fact, the facilitation bilinguals benefit from during childhood may result in a processing "cost" during adulthood. For example, Costa et al. (2000) provided reaction time data indicating that bilinguals were slower than monolinguals in picture naming tasks involving lexical access.

Another possibility is that bilinguals take less advantage of visual speech than monolinguals because they are neurally better "equipped" to process auditory information. Golestani et al. (2007) measured the volume of Heschl's gyrus in French participants who learnt a Hindi dental/retroflex consonant contrast "fast" and "slow." Heschl's gyrus is located in the auditory cortex and is the first cortical area that receives auditory information coming from the peripheral auditory system. They observed that the fast participants had bigger Heschl's gyrus volumes than the slower ones. According to the authors the bigger Heschl's gyrus volumes in fast participants could make them have a better temporal representation of sounds. This would be extremely useful to discriminate the rapid acoustic transitions that we observe in many consonants and thus enhance discrimination abilities for the dental/retroflex Bengali consonant contrast. Furthermore, Ressel et al. (2012) measured monolinguals' (Spanish) and bilinguals' (Spanish-Catalan) volume of Heschl's gyrus. They provided evidence indicating that bilinguals have larger Heschl's gyri than monolinguals. The voxel-based morphometry data for the left Heschl's gyrus indicated that the gray matter volumes were more important in bilinguals than monolinguals. The positive correlation between larger Heschl's gyri and the ability to perceive non-native phonemes suggests that bilinguals would have better auditory capacities to discriminate the Bengali phonemes and would rely less on visual speech. Although this hypothesis is very appealing, it is not supported by our results, since monolinguals and bilinguals had equivalent scores in the audio-only condition for the discrimination of the Bengali dental/retroflex consonant contrast.

The fact that monolinguals were more efficient than bilinguals in the audiovisual condition could also reveal another "bilingual

cost" that may have nothing to do with phoneme identification *per se* but with the visual processing of the speaker's face. To decode visual speech in face-to-face communication we have to process the speaker's face. We have to do a configural analysis to locate the mouth with respect to the eyes, nose, etc. We will then be able to process the movements that transmit the relevant information on phoneme identity. This means that there could be a link between face processing and speech perception. If so, would the bilingual perceptual narrowing pattern observed for visual speech (Sebastian-Gallés et al., 2012) result, or be related to, perceptual narrowing in face processing? From a developmental perspective, face processing and phoneme identification are both important for early communication. Faces can be seen as providing an early channel of communication prior to the onset of gestural or oral language between infant and caretaker (Pascalis et al., 2014). The idea of a link between face processing and lip-reading is not new. In 1986, Bruce and Young's face recognition model already included an optional *facial speech analysis* module that categorized oro-facial movements (Bruce and Young, 1986).

Moreover, some studies suggested that bilingual exposure could lead to changes in brain organization and affect face perception and spatial localization tasks that are linked to hemispheric asymmetry (Sewell and Panou, 1983). More recently, Haussman et al. (2004) investigated hemispheric specialization differences between German monolinguals and German-Turkish bilinguals during linguistic and face-discrimination tasks. The results indicated that bilinguals do not have the same left visual field advantage than monolinguals during face discrimination. Bilinguals' reaction times were longer than monolinguals' when the faces were presented in the left visual field, indicating a difference in cortical organization for face processing between the two populations. These temporal differences are consistent with our study. We also observed that bilinguals' reaction times were slower than monolinguals' in the audiovisual condition. On this basis, and if bilinguals and monolinguals process faces differently, it can have an impact on their abilities to process visual speech. Further research has to be done, of course, to investigate whether the bilinguals' lower phoneme identification scores and higher RTs with respect to monolinguals in the audiovisual presentation could be due to differences in face processing.

To conclude, linguistic experience has an impact on the way we process visual speech. Monolinguals are more accurate and faster than bilinguals to process the speaker's articulatory gestures. This gives them an advantage when having to identify phonemes. Indeed, the former benefit more from visual information than the latter with respect to audio-only communication.

ACKNOWLEDGMENTS

This research was supported by a grant from the Spanish Government, PSI2011-23033 and a grant from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013 Cooperation grant agreement n° 613465 - AThEME).

REFERENCES

Altarriba, A., and Heredia, R. R. (2008). *An Introduction to Bilingualism. Principles and Processes*. New York; London: Lawrence Erlbaum Associates.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Bates, D. (2005), "Fitting linear mixed models in R," in *R News*, 5, 27–30. Available online at: http://CRAN.R-project.org/doc/Rnews/

Bates, D. M., and Maechler, M. (2009). *lme4: Linear Mixed-Effects Models using S4 Classes*. R package version 0.999375-31. Available online at: http://CRAN.R-project.org/package=lme4

Best, C. (1995). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience. Issues in Cross-Language Research*, ed W. Strange (Baltimore, MD: York Press), 171–204.

Best, C. T., McRoberts, G. W., and Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *J. Acoust. Soc. Am.* 109, 775–794. doi: 10.1121/1.1332378

Bruce, V., and Young, A. (1986). Understanding face recognition. *Br. J. Psychol.* 77 (Pt 3):305–327.

Burfin, S., Savariaux, C., Granjon, L., Sanchez, C., Tran, T. T. H., Soto-Faraco, S., et al. (2011). "Overcoming phonological deafness in L2 conversations by perceiving the facial movements of the speaker," in *Workshop on Bilingualism: Neurolinguistic and Psycholinguistic Perspectives*, (Aix-en-Provence), 41.

Burns, T. C., Werker, J. F., and McVie, K. (2003). "Development of phonetic categories in infants raised in bilingual and monolingual environments," in *Proceedings of the 27th Annual Boston University Conference on Language Development*, eds B. Beachley, A. Brown, and F. Conlin (Boston, MA: Cascadilla Press).

Byers-Heinlein, K., Burns, T. F., and Werker, J. F. (2010). The roots of bilingualism in newborns. *Psychol. Sci.* 21, 343–348. doi: 10.1177/0956797609360758

Costa, A., Caramazza, A., and Sebastian-Gallés, N. (2000). The cognate facilitation effect: implications for models of lexical access. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 1283–1296. doi: 10.1037/0278-7393.26.5.1283

Davis, C., and Kim, J. (2001). Repeating and remembering foreign language words: implications for language teaching systems. *Artif. Intell. Rev.* 16, 37–47. doi: 10.1023/A:1011086120667

Golestani, N., Molko, N., Dehaene, S., LeBihan, D., and Pallier, C. (2007). Brain structure predicts the learnin of foreign speech sounds. *Cereb. Cortex* 17, 575–578. doi: 10.1093/cercor/bhk001

Haussman, M., Durmusoglu, G., Yazgan, Y., and Güntürkün, O. (2004). Evidence for reduced hemispheric asymmetries in non-verbal functions in bilinguals. *J. Neurolinguistics* 17, 285–299. doi: 10.1016/S0911-6044(03)00049-6

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., and Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *J. Acoust. Soc. Am.* 119, 1740–1751. doi: 10.1121/1.2166611

Kuhl, P., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Dev. Sci.* 9, F13–F21. doi: 10.1111/j.1467-7687.2006.00468.x

Kuhl, P. K., Tsao, F.-M., and Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9096–9101. doi: 10.1073/pnas.1532872100

Marian, V., Blumenfeld, H., and Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 1–28. doi: 10.1044/1092-4388(2007/067)

Navarra, J., and Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of L2 sounds. *Psychol. Res.* 71, 4–12. doi: 10.1007/s00426-005-0031-5

Pallier, C., Bosch, L., and Sebastián-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition* 64, B9–B17. doi: 10.1016/S0010-0277(97)00030-9

Pascalis, O., Loevenbruck, H., Quinn, P., Kandel, S. Tanaka, J., and Lee, K. (2014). On the linkage between face processing, language processing, and narrowing during development. *Child Dev. Perspect.* 8, 65–70. doi: 10.1111/cdep.12064

Polivanov, E. (1931). La perception des sons d'une langue étrangère. *Change* 3, 111–114.

Pons, F., Lewkowicz, D., Soto-Faraco, S., and Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10598–10602. doi: 10.1073/pnas.0904134106

Ressel, V., Pallier, C., Ventura-Campos, N., Diaz, B., Roessler, A., Avila, C., et al. (2012). An effect of bilingualism on the auditory cortex. *J. Neurosci.* 32, 16597–16601. doi: 10.1523/JNEUROSCI.1996-12.2012

Sebastián-Gallés, N., Albareda-Castellot, B., Weikum, W. M., and Werker, J. F. (2012). A bilingual advantage in visual language discrimination in infancy. *Psychol. Sci.* 23, 994–999. doi: 10.1177/0956797612436817

Sebastián-Gallés, N., and Soto-Faraco, S. (1999). Online processing of native and non-native phonemic contrasts in early bilinguals. *Cognition* 72, 111–123. doi: 10.1016/S0010-0277(99)00024-4

Sewell, D. F., and Panou, L. (1983). Visual Field asymmetries for verbal and dot localization tasks in monolingual and bilingual subjects. *Brain Lang.* 18, 28–34. doi: 10.1016/0093-934X(83)90003-2

Sundara, M., Polka, L., and Genesee, F. (2006). Language-experience facilitates discrimination of /d-ð/ in monolingual and bilingual acquisition of English. *Cognition* 100, 369–388. doi: 10.1016/j.cognition.2005.04.007

Von Holzen, K., and Mani, N. (2012). Language non-selective lexical access in bilingual toddlers. *J. Exp. Child Psychol.* 113, 569–586. doi: 10.1016/j.jecp.2012.08.001

Weikum, W., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., and Werker, J. (2007). Visual language discrimination in infancy. *Science* 316, 1159. doi: 10.1126/science.1137686

Werker, J., and Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev.* 7, 49–63. doi: 10.1016/S0163-6383(84)80022-3

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 March 2014; accepted: 29 September 2014; published online: 21 October 2014.

Citation: Burfin S, Pascalis O, Ruiz Tada E, Costa A, Savariaux C and Kandel S (2014) Bilingualism affects audiovisual phoneme identification. Front. Psychol. 5:1179. doi: 10.3389/fpsyg.2014.01179

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

APPENDICES

Table A1 | Statistical analyses on accuracy values for group (bilinguals, monolinguals), modality (A-only, AV), and phoneme type (native, non-native).

Factors	*t*-value	Pr(>|*t*|)
Group	−1.073	0.2835
Modality	2.343	0.0192
Phoneme	−0.603	0.5465
Group*Modality	3.287	0.001
Group*Phoneme	2.286	0.0223
Modality*Phoneme	3.215	0.0013
Group*Modality*Phoneme	−1.143	0.253

*Indicates interaction.

Table A2 | Statistical analyses on Reaction time values for group (bilinguals, monolinguals), modality (A-only, AV), and phoneme type (native, non-native).

	t-value	Pr(>|*t*|)
Group	−3.33	0.0009
Modality	−2.34	0.0191
Phoneme	−3.03	0.0024
Group*Modality	−3.02	0.0025
Group*Phoneme	1.15	0.2489
Modality*Phoneme	−0.36	0.716
Group*Modality*Phoneme	0.87	0.386

*Indicates interaction.

RESUME

En situation de perception audiovisuelle de la parole, comme lors des conversations face-à-face, nous pouvons tirer partie des informations visuelles fournies par les mouvements oro-faciaux du locuteur. Ceci améliore l'intelligibilité du discours. L'objectif de ce travail était de déterminer si ce « bénéfice audiovisuel » permet de mieux identifier les phonèmes qui n'existent pas dans notre langue. Nos résultats révèlent que l'utilisation de l'information visuelle permet de surmonter les difficultés posées par la surdité phonologique dont nous sommes victimes lors d'une présentation auditive seule (Etude 1). Une étude EEG indique que l'apport des informations visuelles au processus d'identification de phonèmes non natifs pourrait être dû à une modulation précoce des traitements effectués par le cortex auditif primaire (Etude 2). En présentation audiovisuelle les phonèmes non natifs donnent lieu à une P50, ce qui n'est pas observé pour les phonèmes natifs. Il semblerait également que l'expérience linguistique affecte l'utilisation des informations visuelles puisque des bilingues précoces semblent moins aptes à exploiter ces indices pour distinguer des phonèmes qui ne leur sont pas familiers (Etude 3). Enfin, l'étude de l'identification de consonnes plosives natives avec une tâche de dévoilement progressif nous a permis d'évaluer la contribution conjointe et séparée des informations auditives et visuelles (Etude 4). Nous avons observé que l'apport de la modalité visuelle n'est pas systématique et que la prédictibilité de l'identité du phonème dépend de la saillance visuelle des mouvements articulatoires du locuteur.

Mots-clés : perception audiovisuelle de la parole, phonèmes natifs et non natifs, bilinguisme, neurophysiologie.

ABSTRACT

During audiovisual speech perception, like in face-to-face conversations, we can take advantage of the visual information conveyed by the speaker's oro-facial gestures. This enhances the intelligibility of the utterance. The aim of this work was to determine whether this "audiovisual benefit" can improve the identification of phonemes that do not exist in our mother tongue. Our results revealed that the visual information contributes to overcome the phonological deafness phenomenon we experience in an audio only situation (Study 1). An ERP study indicates that this benefit could be due to the modulation of early processing in the primary auditory cortex (Study 2). The audiovisual presentation of non native phonemes generates a P50 that is not observed for native phonemes. The linguistic background affects the way we use visual information. Early bilinguals take less advantage of the visual cues during the processing of unfamiliar phonemes (Study 3). We examined the identification processes of native plosive consonants with a gating paradigm to evaluate the differential contribution of auditory and visual cues across time (Study 4). We observed that the audiovisual benefit is not systematic. Phoneme predictability depends on the visual saliency of the articulatory movements of the speaker.

Key words: audiovisual speech perception, native and non native phonemes, bilingualism, neurophysiology